1.1 Statistics in biological systems

In the last decades we have witnessed an exponential growth of two big forces driving our knowledge enrichment: computational efficiency and big data. While the former encompasses a broader range of particulars - raw speed improvement, parallel computing, GPU processing - when compared to the latter, the sole term big data has rapidly capitalized all the attention of statistical community (among others), even if being able to precisely tell how big is ‘big’ - in some situations - has been proven to be quite difficult. It is worth noticing that neither of them, however, when purely approached without careful reasoning, can guarantee a clearer insight about complex systems that naturally describe interesting phenomena, such as those arising in biological contexts. Sure enough, when faced with a complex task, the first - and strongest - instinct has dangerously shifted towards throwing more and more data to the model, backed-up by both the vast amount of information available and the improved computational capacity of handling it. The first issue is that under the dome of big data most of the traditional - in the broadest meaning of the word - statistical techniques and methodologies lose their feasibility, both for practical and theoretical reasons. A perilous behavior threatening statisticians nowadays is to rely on complicated models to answer (simple, albeit not easy) questions about complex systems, calling on ‘bigger’ dataset for support on their assumptions and beliefs. In an attempt to balance out complexity, interpretability of the results and feasibility of a statistical analysis, a reasonable and careful step would be instead to start approaching the problem by modelling the structures under-
1.1 Statistics in biological systems

lying the data at our disposal. Effectively, knowledge may very well not be strictly equal to perfect description of a phenomenon but, instead, conceived as the understanding of the relationships between its fundamental components (be them genes, proteins, or generic statistical units per se). As an example, we may think about the complexity of some biological processes: gene-gene or gene-protein interactions, cellular signaling, circadian clock cycles, molecules dynamics and so forth. Especially in this fascinating contexts, along with - for example - social networks, we have been flooded with information: faster and cheaper technologies to biochemically analyze genetic data, labelled under the term next-generation sequencing (NGS), are introduced with increasing pace. The number of observations (i.e. the number of statistical units) and ‘features’ observable in a single experiment are skyrocketing. What kind of dependencies should we incorporate in our model, it being a simplified version of such a complex reality? A preliminary answer to this question, when thinking about the description of a process governed by stochasticity, is the spatio-temporal dimension. As natural as it may seem, however, without a solid a priori idea of the underlying process we might be forced to forego this ‘quick and safe’ structure given by time and space: that is, the strongest focus needed at the beginning is about the ever present trade-off between the idea of exploring data and looking for confirmatory results. If we want to better understand the spatial dynamics of interactions between genetic markers, for example, it might be sensible to encode a rich and flexible dependency structure in the model while keeping a simple level of description of the actual measurements comprising the dataset. On the other hand, if we are trying to analyze specific aspects of a process - again, for example, gene interactions - it is a powerful approach to encode in the model as much prior knowledge we have about the underlying dependency structure governing the units’ relationships, while aiming for a richer description of the observable level of the model. The main flavor about this idea is to build a model in a hierarchical fashion, with a distinction between unobservable (latent) quantities and observable information, and to fit our idea of the spatio-temporal dimension where it truly belongs with respect to this two layers; the modularity of the aforementioned scheme gives also rise to another encouraging aspect: to seek extensions of the model proposed,
updating step-by-step the levels of the hierarchy. Out of the possible paradigms to follow, we will embrace the Bayesian one, mostly for two reasons: the built-in quantification of uncertainty of the parameters in the model and the more straightforward possibility to tackle a complex model. We will also switch from the epitome of big data to high-dimensional data, which bears a more focused phrasing on what could be the associated characteristics of this kind of data (curse of dimensionality, sparsity, number of features higher than available samples and so forth).

1.2 Thesis outline

We will outline a path to follow while exploring the aforementioned challenges, with a focus on two main tasks: clustering genes and retrieving information about unobservable quantities. While introducing the spatio-temporal structure, we will move from considering both time and space dependencies, but in a simple fashion, to model them separately yet in a richer way.

The PhD project is structured in three parts. In the next chapter we will present a paper that employs a Markov Random Field (that is, a graphical model) as a latent structure to describe the spatial relationships between locations of the genome and the temporal dependency of experimental replicates of the very same strand of DNA; faced with the task to understand if a location is configured as enriched or not by a protein, we will use a mixture of two discrete distribution to describe the observed counts of a particular genetic measurement, while simultaneously dealing with specific issues intrinsic to the NGS context.

In the third chapter, a non-conventional model-based clustering scheme will be presented, that allows for a unit (a gene, in our case) to belong to more than one cluster at a time. In allocating these units, however, we will take into account the natural neighboring spatial dependency between them and we will model it so that it plays a role into the in the way we cluster our observations through the weights of the chosen mixture of discrete distributions.

In the forth chapter, we will propose a Bayesian hierarchical model for Ordinary Differential Equations (ODE) describing the temporal dynamics of a continuous process: while searching for a flexible description of the data through the use of penalized spline regression, we will indirectly solve the ODE and quantify the uncertainty about the so-
solution obtained with respect to the noisy observations we have at our disposal.

1.3 Chapter 2: Mixtures and graphs

In the second chapter of this thesis, we first introduce and model data coming from chromatin immunoprecipitation and sequencing (ChIP-Seq) experiments. Whereas years ago microarrays were considered ‘gold standards’ for genetic related analysis, Next-Generation Sequencing has now become the prominent broad class of biological techniques employed to study the complex relationships between the DNA, RNA, proteins and cell functionality. In this context, with ChIP-Seq experiments, we are interested in discovering if a protein of interest is binding or not to the RNA and to which portion of it. More specifically, if a region has an associated ‘high’ count in the data it is more likely the protein has enriched (i.e., bound to) that region. If we know the aforementioned protein to be associated with a particular disease and we find out which portions of the DNA of a cell, from an affected tissue, the protein is enriching we may infer that those regions - and, more in depth, the individual genes - are associated with the disease itself. From a statistical point of view, a not to be overlooked feature that sets ChIP-Seq (and in general NGS experiments) apart from microarray data is that they yield discrete observations instead of continuous measurements. Other peculiarities are: overdispersion, that is an amount of variability greater than expected when a basic model (like the Poisson distribution) for the counts is assumed; even though usually raw data are pre-processed, the units studied relate to region of the DNA that are contiguous to each other and thus a need for a spatial pattern to be included in the model arises; zero-inflation, that is an abundance of zeros as observed values. Also, as efficiency progresses and costs decrease, more biological or technical replicates are made available to the researchers from the laboratory, demanding for statistical methodologies able to consider jointly all the information at their disposal. Last but not least, the measurements are collected at different time points throughout the whole experiment, giving the opportunity to look for a temporal dependency in the data. To face these challenges, we propose a hierarchical mixture model that fuses together two distinct layers: a latent structure, devoted to infer the unobserved protein binding process we are interested in, built by considering both the spatial and tem-
poral dependencies that connect the regions of the DNA analyzed in the experiment; and secondly, a measurement model, characterized by means of a mixture of discrete distributions that can accommodate for overdispersion and jointly model all the technical or biological replicates. For the hidden layer, we choose a Markov Random Field, to reflect the assumption of first order spatio-temporal dependency: that is, each region (unit) depends only on its left and right contiguous neighbors - in a spatial sense - and previous or following neighboring time point - in a temporal sense. The latent structure can be represented through a graphical model and thus all the related theory can be used to translate it into probabilities. As for the mixture model, we consider Negative Binomial distributions for their flexibility and innate description of overdispersion. Following a Bayesian approach and all the standard derivations needed to implement an MCMC procedure, we first assess the performance of our proposed method in a simulation environment and then we move to the analysis of differential roles in gene regulation of two transcription factors, p300 and CBP by means of ChIP-Seq data.

1.4 Chapter 3: Spatial dependency and beyond

In the third chapter of the thesis, we look at the same experimental framework, ChIP-Seq data, but from a different perspective. We are now interested in classifying the units (genes, regions of the chromosome, etc.), based on the observed counts, into groups (clusters) that have a meaningful biological interpretation. What we have in mind about "meaningful biological interpretation" is mostly related to a simple paradigm: cell functionalities (for example mitosis, self-destruction, stress response) are regulated with complex patterns of signaling by genes, proteins and other important biological ‘actors’. It is not unlikely that a gene, a protein, a biological ‘actor’, participate in more than one cell functionality. If we now translate "gene, protein, actor" and "cell functionality" into a more statistical language as "units, observations" and "clusters" we can then interpret the original problem from a more methodological point of view. Model-based clustering is a widely used technique that has found application in a broad class of problems in many different contexts. However, in his simplest description, the framework assumes the groups are mutually exclusive and this is a restriction that does not fit the motivating re-
search question. In the literature, some authors proposed extension to overcome this limitation; having in mind the same challenges and peculiarity of the data that we already described in Section 1.3, we propose a model that can accommodate at best all of them: allocation of units into - potentially - more than one group; discrete measurements; overdispersion; clusters that are interpretable as "meaningfully" linked to each other (where, again, meaningful is a term recalling the cell functionality paradigm). We pursue (as before) the solution of the task by employing a hierarchical structure. We use a mixture of Negative Binomials to flexibly model the observations and a latent allocation structure that classifies each unit into groups. What we define as groups can either be: ‘primary clusters’, which represent the principal functionalities or prototypical functions that we have in mind in our research question (for example, cluster of enriched regions and cluster of non-enriched regions); ‘multiple allocation clusters’, that are collection of units that belong simultaneously to more than one of the primary clusters. The main idea is to move from the original representation of a mixture model for clustering with, let’s say, $k = 3$ groups to a mixture that has $2^k = 8$ groups, where $2^k = 8$ is the number of all possible configurations of allocation we could have for one unit. In doing so, we create multiple allocation components (clusters) having parameters that are completely specified from the primary clusters originating them. In this case, thus, we are actually not introducing in the model new parameters that have to be estimated. As a by-product of this approach, we obtain an ‘outward’ cluster: a group in which outliers could be allocated or, as we do, observations that would otherwise be account for with a zero-inflation adjustment of the model. A further step that we take is introducing the spatial dependency among the units we are studying. We do so by allowing the weights of the mixture, which reflect the prior probability to be allocated to a cluster, to vary according to a spatial pattern that we assume to be a conditional autoregressive model. More specifically, this pattern takes into account the position and the relative distances between the units and encodes them as an additional latent layer into our hierarchical structure. To study the performance of the proposed model we test it in a simulation environment and then we proceed to inspect the same dataset as in Chapter 2 from this alternative point of view.
1.5 Chapter 4: Time dynamics and complex systems

In the fourth chapter, we focus on the dynamics of biological systems that evolve through time. In many fields of application, such as engineering and the study of dynamical systems in biology, chemistry and physics, researchers often describe the behavior of complex systems with a set of equations called ordinary differential equations (ODE’s). These mathematical objects are equations that attempt to model the changes of the state of the system with respect to time by considering a set of parameters, unknown quantities governing the law of the process itself. There is a dual aspect to be considered when looking at the system of ODE’s: the known relationships are expressed at the derivative level of the components of the process but what we actually observe are the states at fixed time points, not the evolution in time. Even if the functionals in the equations are known, expressing the dependency among the components, we still need to estimate the parameters in order to fully describe the dynamics. If an analytical solution of the system is available, those parameters can be directly recovered from the observed data; if such a solution is not readily available in closed form, as it is often the case, numerical integration is needed. In many contexts it is also likely that the data we collect are affected by noise, perturbing the real underlying dynamic. Another way to look at the task at hand, a perspective more deeply connected with the statistical methodology, has been followed in the literature in the last decade: it is possible to avoid direct integration of the equations by smoothing the data and many methods have been proposed, either within a frequentist and Bayesian framework. We follow this aforementioned idea by proposing a two-step Bayesian Smooth-and-Match strategy: at a first stage, we smooth the observed data in order to reconstruct a noiseless sequence of the states of the system for all its components; at a second stage, we use this states as inputs for the known functionals that appear in the ODE’s and we ‘match’ this temporary description of the dynamic with the data we have. We thus move the focus from solving the system to directly infer the parameters that describe the process, obtaining the solution as a by-product of all the procedure. Delving briefly into some technicalities: we adopt penalized Bayesian smoothing with cubic splines as our first step of the procedure; as for the second step, we obtain the parameters of the
system through a penalized (ridge) regression approach. The two compartments of the strategy are finally connected by assuming a common noise term that acts as a built-in quantification of the uncertainty associated to the solution of the system that we have (again, indirectly) reconstructed within the MCMC sampling scheme. In order to assess the performance of the proposed method we set up three different simulation studies and we then proceed to compare the results we obtain on another dataset previously analyzed by other authors.