Big data analytics in bioinformatics

9 Valorization

The continuous production and accumulation of biological and biomedical data have led the scientific community into a "big data era". In this era, data and computers have become one of our first-line defenses against disease development and progression. With the vast amount of data comes new opportunities and possibilities, which have motivated the construction of novel powerful analytical and machine learning algorithms that can locate and extract key patterns from the various and large data sets. Already, the marriage of big data and machine learning is used for diagnostic and prognostic purposes in clinical settings, and to identify new drug targets and purposes of validated drugs. Thus, the continuation of efficient data handling and the development of algorithmic analytical models are vital for the perpetual battle against diseases, disorders and illnesses.

Data mining and algorithmic analyses of biological and biomedical data is typically carried out by bioinformaticians, who combine computer science, mathematics, statistics and biological and biomedical knowledge to discover new insights and previously hidden patterns. It has now become evident that bioinformaticians are playing and will play a key role in carrying the weight of the opportunities, possibilities and responsibilities that have emerged with the large amounts of data produced in recent years. Now more than ever is the collaboration amongst and across scientific fields essential, as the bioinformaticians' classifiers, clustering algorithms and simulations always need guidance from biologists and medical staff whose expert knowledge serves as important feedback for the ongoing endeavor of improving the data driven models and their predictive capabilities. Such (direct and indirect) collaborations are in fact part of an already existing loop that (roughly) goes from biological data → bioinformatic models and predictions → "wet lab" experiments → biological data → … and so on. This loop ensures that both experimental data and bioinformatic solutions keep refining each other, which synergistically pave the way for new insights, discoveries and innovations.

Alongside the above-described "knowledge-loop" that strongly influence the design of current biomedical hypotheses and approaches, an interesting and significant scientific movement is beginning to pick up momentum. The expanding terrain of new biomedical data and novel algorithmic approaches is enabling researchers to see biological phenomena in a new and more clear light and to find more detailed explanations of observed molecular events. A pivotal example of this tendency is the changing conceptual understanding of diseases and their definitions. The current disease definition is one based on symptoms, however, with the growing amount of data, knowledge and insights scientists and clinicians are beginning to dig deeper for satisfying explanations and nosological connections. And it is getting clearer that diseases are more meaningfully defined in terms of causal mechanisms instead of the symptoms they generate, as the same symptoms can be produced by distinct disease-causing mechanisms. A bonus of such a redefinition of the disease term is that the specific and underlying cause is already identified, which again

readily reveals how to treat it via e.g. personalized medicine and network pharmacology approaches. Thus, bioinformatic work is an important link in the chain of our scientific evolution, as the bioinformatic solutions are key factors in driving the progress and for keeping the momentum of the slowly but steadily accelerating scientific movement.

This thesis presented three manuscripts that each either indirectly or directly showed how a synergetic use of bioinformatic solutions and experimental data can lift the impact of scientific discoveries to new heights. Additionally, all the manuscripts revealed findings that can aid a general mechanism-based disease understanding and therapy development.

Starting with the first included manuscript (chapter 2), here we presented a new clustering method for scRNA-seq data that can analyze and compare single-cell development trajectories. For data from studies on healthy cell differentiations and disease progression, we showed that our innovative method can identify mechanistic patterns involved in driving the developments of the cells. We hypothesize that our tool's findings can serve as key steps in drug repurposing pipelines and perhaps help formulate mechanism-based disease definitions. For instance, when we analyzed the scRNA-seq data from a study on CD8 T-cell development in chronic infections, we located potential drug targets that may help establish new hypotheses about therapy development.

In the second manuscript (chapter 3), we introduced a new simple neural network classifier and demonstrated that the network can reveal how nucleotide variants affect binding affinity of RNA-binding proteins. The predictions and binding profiles produced by our tool directly relates to the discovery of mechanisms underlying diseases. And as a new feature, we showed how our models can be used to guide the development of SSO-therapies (splice-switching oligonucleotide therapies), which might be a valuable technique in clinical settings and for the medicinal industry. A known disease that is treated with SSO-therapy is Spinal Muscular Atrophy (SMA). We are convinced that it is only a matter of time before many similar treatment approaches will be invented and pass the clinical trials.

The third manuscript (chapter 4) described a simulation approach that was invented with the sole purpose of unraveling mechanisms underlying a detected PKG-dependent downregulation of ROS-producing proteins. And as it was suggested in the manuscript, the discovery of such a mechanism may be crucial to the progress of diagnosis and treatments for ischemic stroke induced damages. Both PKG and ROS signaling are represented by drugs in clinical practice. Several more applications are possible and being tested; some have already failed. However, it is highly likely that this was due to the fact that patients were included solely based on clinical parameters and not on mechanistic ones. Precision medicine needs both precision diagnostics and precision mechanism-based drugs. With the here generated tools, the search for new mechanisms and mechanism-based diagnostics

may be furthered. Nevertheless, further experimental analyses and clinical validation are needed before anything can be refined and/or concluded.

As it can be seen, this thesis is a child of its “scientific time”, as it addresses present and ongoing challenges, and is interlinked with the general and established knowledge-refining feedback loop and the scientific movement that the big data era has helped to bring about. Altogether, this thesis bears relevance for biology, biomedicine, bioinformatics and treatment development.