Editorial: Accelerating systems biology

Over the years, the long-standing questions concerning disease evolution, plausible cellular behavior and the hidden properties of organisms have become more and more pressing. They sparked a movement of experts from an array of scientific areas and stimulated them to work cooperatively with the ambitious aim of deciphering the language of nature. From the onset, the field of genomics has been built on many important discoveries beginning with the DNA double helix structure in the 1950s, reverse transcriptase in the late 1960s, recombinant DNA and restriction enzymes in the 1970s, and finally, the polymerase chain reaction discovery in the early 1980s. For some time the polymerase chain reaction was particularly revolutionary because it was the only method used to determine the base sequence of DNA. Wishing to understand how key components ‘converse’ in time, space and across multiple organizational levels became one of the driving factors that led scientists to sequence large eukaryotic genomes, such as that of human, and to the discovery of more than 20,000 different genes and of nearly a million proteins within the cell. The transition years between the 1990s and first decade of the 2000s were the time of the collection of such results that were later mathematically formalized and summarized in artificial wired diagrams.

Soon after, the idea that the full understanding could be possible by a holistic rather than by an atomistic approach became slowly appealing. Investigation on the assembly of the cell’s elementary parts to form complex structures represented the main activity of a promising new research field named ‘Systems Biology’.

Progress in this area required interdisciplinary breakthroughs and close links between wet and dry experimentations. Due to an enormous eager interest, synthetic formalizations (or models) grew rapidly in number and size, models that later could be stored in resource databases easily accessible to anyone of interest. ‘BioModels.Net’ is one of such data resources. It is a database of models ‘curated’ by Le Novère and co-workers. In their paper, they offer an overview of some peculiar characteristics of the data bank in light of the computerized web services they provide to ease the user impact.

In spite of this, models became unmanageable by manual inspection because of their size and compelled biologists to reduce drastically the kind and quality of the achievable analysis procedures. Those whose job was that of modeling living systems very soon acknowledged how badly the effectiveness of such procedures is undermined by the complexity of the resulting models, which, in fact, are a reflection of the complexity of nature itself. If computer-driven experimentation had not been introduced almost alongside, any outcome would have been incomplete. Many automatic analysis procedures were made available to the scientific community, such as steady state, bifurcation, robustness analyses, model checking, simulation, etc. With regard to this, ‘CaliBayes’ and ‘BASIS’ represent the efforts made by Wilkinson and co-authors to collect and integrate some tools for calibration, simulation and storage of biological simulation models. Alongside them, Uhrmacher’s group has also proposed a valuable modular framework, ‘James II’. It works to ease the process of designing wetlab experiments through advanced routines and algorithms. Both the goals match: they aim at providing new insights for new models and to feed back the so-called ‘hypothesis-driven science cycle’.

Dry laboratories proliferated rapidly. To better describe their biological dynamics of interest from precise perspectives, each defined a new appropriate language. As a first result, this gave rise to an explosion of new software tools along with an overall ‘Babel of voices’, since each one of them talked a different dialect. The need for a common language was becoming a ‘must’. Standard languages (like SBML and CellML) were created with the primary aim of making cooperation among software tools possible. They were able to bridge the gap between wet and dry scientists by broadening the range of the available computational methods. They definitely reached their goal.
In the case of modeling, successive developments focused on how to deal with more and more complex systems. Hence, increasing evidence in the constant interplay between components of different biological systems (cross-talk) proved that isolated models are quite unrealistic and that any analysis result is imperfect. Therefore, sets of artificial models of higher chemical density and complexity were merged and analyzed, however, without any significant results. In fact, the exponential growth of the computational power required to deal with huge ‘state-spaces’. This had the simple effect of hindering any of the aforementioned methodologies, and thus causing them to ultimately flop. For this reason Safranek, Brim and Barnat have developed ‘DIVinE’, a model checker made up of a collection of tools designed to exploit the capabilities of the new hardware architectures, as well as to restrain the state space of complex models from exploding.

One of the main limitations in managing biological models comes from the fundamental difference between evident high parallelism in biochemical reactions and sequential environments employed for the analysis of these reactions. Such limitations affect all varieties of continuous, deterministic, discrete and stochastic models by undermining the applicability of simulation techniques and the analysis of biological models.

Parallel and distributed computing may be deployed to compensate for this lack. They rely on both the intrinsic parallelism of nature and the power of multiprocessing architectures. Indeed, in real life any biological aspect is (to some extent) parallel. If a chemical transformation occurs, it does not take place for two molecules, but, as a principle, for all molecules. In a less gross view, hundreds of independent biological transformations take place simultaneously rather than in a sequential manner. This shows that natural phenomena can be seen as massively parallel processes, since they occur above at least two independent levels of parallelism. The work by Cecilia and colleagues provides an example of a formal language that highlights the massive parallelism typical of biology. Based on this language and on CUDA, the computing engine running in NVIDIA graphics processing units, they present their software setting.

Revolutionary biology demanded revolutionary computing. Great goals have been achieved since the SPARCcenter 2000 was used to assemble the genome of *Haemophilus influenzae* in 1995. That was the last model of a computer generation limited by an architecture capable of addressing only 2 GB of RAM. Modern processors are almost 200 times more powerful. The cost of storage has also dropped dramatically over the years. In 1992, 1 TB of disk space cost 1 million dollars; in 2010, the cost has been reduced to nearly 100 dollars. Furthermore, while in 1992 10 MB/s was the typical speed in networks, today gigabit network interfaces are very common and specialized processors are affordable for all.

Prandi and Dematte provide an overview of how Systems Biology has been affected by such circuits as a whole. In this direction, works of Richmond, Walker, Coakley and Romano and that of Bako are particularly noteworthy. The former makes use of the flexible large-scale agent modeling environment called ‘FLAME’ to simulate cellular-level processes on graphics processing units, while the latter demonstrates that artificial spiking neural networks built to resemble the biological model encoding information in the timing of single spikes are capable of computing and learning clusters from realistic data making use of embedded soft-core microcontrollers.

Thus, the very nature of large-scale computing has changed from systems relying on one or a few powerful custom-designed processors to scalable parallel systems or farms of computer. Any automatic procedure is approaching an alternative parallel implementation; and several, not trivial, ad hoc synchronization policies are sometimes coming up to support their final deployment. The very revolution lies in the fact that complex problems can now be broken down into sets of smaller jobs and can be executed on high-performance machines nodes, even remotely available, through easy-to-use access interfaces. The paper entitled *Estimating the divisibility of complex biological networks by sparseness indices* is an exemplary attempt to find a rationale behind the exhausting quest for ‘smaller and clever’ computational subunits.

Hence, to cut a long story short, I am of the view that there are many-sided horizons where researchers are going. If I were to mirror their needs and dreams, I believe they can be reached only by crossing the
rivers of accessibility, feasibility, quickness and reliability, important properties of any future framework. This issue tries to provide the reader with some clue in finding their own way. I hope you will enjoy reading about this and that this issue has done the job of exciting your interest and has been particularly helpful in encouraging you to find a common and long-run target on this path.

Tommaso Mazza
The Microsoft Research – University of Trento
Centre for Computational and Systems Biology
Povo, Trento, 38123, Italy