



Introduction to the Special Theme

Tackling Big Data in the Life Sciences

by the guest editors Roeland Merks and Marie-France Sagot

The Life Sciences are traditionally a descriptive science, in which both data collection and data analysis both play a central role. The latest decennia have seen major technical advances, which have made it possible to collect biological data at an unprecedented scale. Even more than the speed at which new data are acquired, the very complexity of what they represent makes it particularly difficult to make sense of them. Ultimately, biological data science should further the understanding of biological mechanisms and yield useful predictions, to improve individual health care or public health or to predict useful environmental interferences.

Biologists routinely measure DNA sequences, gene expression, metabolic profiles, and protein levels, infer molecular structures, visualize the positions and shape of cells in developing organs and whole embryos over time, rapidly screen the phenotypic effects of gene mutations, or they monitor the positions of species and individual animals and plants in whole ecosystems or herds, to name just a few.

The resulting datasets yield new insight into biological problems, while at the same time they pose new challenges to mathematicians and informaticians. How do we store these large bodies of data in an efficient manner and make sure they remain accessible in the future, while at the same time preserving privacy? How do we search through the data, spot interesting patterns and extract the biologically relevant information? How do we compare data of different conditions or species? How do we integrate data from different sources and across biological levels of organization to make new predictions? Can we then use the resulting patterns to solve the inverse problem and derive meaningful dynamical mathematical models from kinetic datasets? How do we model complex and/or large biological systems, e.g., whole cells, embryos, plants or ecosystems? What are the challenges to multiscale modeling?

Given the great variety of topics that such general questions cover, this Special Theme of ERCIM News could only highlight a few, in a selection of nineteen papers. These provide an overview of the latest techniques of data acquisition at the molecular level, address problems of data standardisation and integration, and propose sophisticated algorithms for analysing complex genetic features. Such algorithms help to infer the rules for predictive, dynamical modelling, by characterising genetic interactions, as well as molecular and cellular structures from noisy and incomplete data. They also provide key data for modelling multiscale systems at the multicellular or ecosystem level. Alongside data mining approaches, such dynamical models are a useful aid for proposing new control strategies, e.g., in personalised medicine or to help improve medical diagnostics. A further selection of articles discusses visualisation methods for big and noisy data, or suggest to make use of new communication techniques, such as Twitter, to help in quick management of health. Overall these approaches illustrate the breadth and the beauty of the mathematical and computational approaches that have been developed to cope with the challenges that biology poses to data science. These challenges will only continue to grow in the foreseeable future, as the volume, quality and types of biological data keep on expanding.

Please contact:

Roeland Merks, CWI, The Netherlands
Roeland.Merks@cwi.nl

Marie-France Sagot, Inria, France
Marie-France.Sagot@inria.fr