WORKSHOP REPORT

Computational Modelling of Complex Spatial Structures and Processes in Natural and Life Sciences

BRUSSELS, 10 DECEMBER 2014

SCIENCE EUROPE
Life, Environmental and Geo Sciences Committee

Workshop on Computational Modelling of Complex Spatial Structures and Processes in Natural and Life Sciences

Introduction

In recent years scientific research has changed markedly. High throughput technologies have resulted in the generation and accumulation of unprecedented quantities of data, which can be collected, analysed and interpreted through advances in computational science. This, in turn, has led to rapid progress in the development of computational models that can represent natural phenomena across the range of scales, from the level of atoms and molecules to the structure and evolution of galaxies.

Such models allow researchers to test and develop new ideas; they can provide new insights into the behaviour of systems under different conditions that would be difficult to test directly through experiment, and they can be used to predict how a system will react to changing circumstances.

With computer modelling becoming an increasingly important tool for researchers, the Science Europe Scientific Committee for Life, Environmental and Geo Sciences organised a workshop to explore how different scientific disciplines use computational modelling and simulations of complex structures in space and time.

Workshop Aims

The specific aims of the workshop were to:

- Discuss current approaches and areas of recent progress (particularly in terms of methodology) across different scientific fields and across different scales, ranging from the level of atoms through to molecules, cells, organisms, populations, environments, the Earth, and up to systems on the astronomical scale, as well as other structured systems that are studied across time scales;

- Identify emerging trends and areas of potential co-operation and synergy by:
  - Assessing the role of knowledge-based techniques, in particular machine learning, in the development of models;
  - Discussing the type and quality of input data that support modelling and reproducibility in scientific research;
  - Comparing methodologies, infrastructures and resources used by different modelling communities;
  - Identifying approaches that contribute to increased reproducibility in scientific research; and
  - Discussing career opportunities and education and training requirements for modelling scientists.

- Formulate recommendations for future actions; and

- Discuss possible follow-up activities (such as workshops and scientific conferences on topics identified as being of particular interest).

Workshop Format

Participants presented ten case studies of how modelling has been used in a range of different scientific contexts, and emerging challenges and opportunities were discussed in the context of these. These case studies were followed by discussions dedicated to emerging trends, synergies, and recommendations.

Case Studies that Demonstrate how Modelling is Providing New Insights into Science

1. Modelling Chemical Kinetics

Chemical engineers need to understand how fundamental chemical processes can be scaled up to produce the optimal design of chemical reactors and guidelines for materials design. Kinetic information derived at the nanometre level in the laboratory needs to reach the metre scale in the chemical reactor. Experimental data combined with theoretical calculations result in an understanding of chemical kinetics at the microscale. Laws of conservation of mass, energy and momentum can then be applied, taking into account issues of mixing, dispersion and transport phenomena. Models of the reaction can then be derived and manipulated to produce the optimal reactor design.

2. Multi-scale Simulations of Biomolecular Systems

Biological systems operate across a range of scales, from electrons and atoms through to molecules and whole cells. Integrating these scales to obtain meaningful simulations of a biological system is a major challenge. Multi-scale modelling attempts to reconcile these scales through the application of a range of parameterisation strategies, both top-down and bottom-up, combining fine- and coarse-grained models across the range of timescales, and integrating this information with experimental results.

3. Modelling of RNA 3D Structures

For many large biological molecules it is the three-dimensional structure that defines the molecule's activity and how it interacts with other molecules. It is becoming clear that RNA has a much more important 'executive' function in the cell than was previously thought, for example compared to proteins. New computational models are being developed to understand how RNA's 3D structure is related to its function. Many of the models are inspired by previously-developed and practically-validated methods for predicting protein structure and function. Models typically investigate either the physical forces that are involved in very fast spontaneous folding of the macromolecule to produce the most energetically favourable 3D conformation, or the slow process of mutation that gradually changes the molecule's sequence and selection that retains structures exerting a preferable function.

4. Molecular-level Modelling of Biological Processes

Within living cells, proteins and other macromolecules interact to create complex biochemical machines. Modelling the structures and interactions of these molecules provides insights into key biochemical pathways and networks within the cell. Highly-detailed structural information may be available for some proteins, while much lower resolution information may be available for others. A central challenge is how to reconcile these different levels of resolution to produce meaningful models of interactions. The aim is to combine theoretical and experimental data on thermodynamics, kinetics and affinities with structural data to produce predictive models of interactions. Furthermore, these networks of interactions need to be framed – and described – in a cellular context that takes into account the effects of the cellular environment, including factors such as pH, ionic strength, redox properties, and the presence of multiple components. Finally, the factors determining the functional process(es) performed by the biomolecule complex or machine need to be rationalised.

5. Systems Biology of Cancer

Researchers studying cancer are able to obtain increasing amounts of information relating to genes and biomolecules in the cell. By taking a series of snapshots over time and under different conditions, changes can be monitored. A fundamental goal is to understand the sequence of events by which a mutation in a gene eventually manifests itself as disease: which proteins and other biomolecules are involved and how they interact in pathways within the cell. By using the data obtained from the snapshots, researchers are seeking to develop computational maps of the biochemical pathways involved in cellular processes. In this way it may be possible to predict how changes in the genome cause disease, opening the way to developing new and more effective therapies.

6. Modelling the Immune System

The human immune system consists of a range of different types of immune cell with distinct and discrete functions that interact in complex networks to protect against invasion by pathogens. The system is complicated and plastic, and many gaps remain in our knowledge. New high throughput cell-counting technologies are yielding information about the size of the populations of different types of immune cell in a given circumstance. This can be integrated with other information, for example gene expression in the cell or the presence of other biomolecules such as proteins, into a series of 'modules' of functionality. This information can, in turn, be used to develop computational models of how these modules interact and are regulated, and how immune cells differentiate.

7. Plant Systems Biology

Polyploidy is a phenomenon that occurs in various plant cells, where DNA is replicated in the absence of cell division, resulting in multiple chromosomes accumulating in a single cell. Polyploidy is important in determining factors such as cell size, differentiation and response to stress. Researchers are developing models of the pattern of polyploidy in the growing root of the plant Arabidopsis. This involves integrating large amounts of data, including gene expression profiles relating to cells with different ploidy levels, which gives information about the spatial distribution of cells of varying ploidy. Predictive maps of ploidy boundaries have been created, validated by experimental methods. These models will enable predictions of how ploidy patterns are induced by stress.

8. Modelling Forest Ecosystems

Models are an important route to understanding a large ecosystem such as a forest because these are often too big to be amenable to direct measurement and changes occur over very long timescales. Understanding forest ecosystems has become important in recent years because of the role of forests in climate change and the impact of climate change on forests. Models are being developed that, for example, simulate the dynamics of carbon, water and nitrogen within forest ecosystems, and the ability of pine forests to remove particulate matter from the atmosphere.

9. Global Nutrient Cycles and the Earth System

Highly-complex computer models have been built, modified and refined over many years to help understand the role that carbon dioxide plays in climate change and to predict future potential scenarios. An important question is whether there are any key components that are missing in these models. Some of the models take little account of global nutrient cycles, such as those for nitrogen and phosphorus. These nutrients are essential for the growth of plants and micro-organisms which, in turn, have a vital role in the carbon cycle. Computer models of nutrient cycles may therefore become, in the future, more important components of Earth climate models.

10. Modelling the Structure and Evolution of our Galaxy

To understand the structure of the galaxy and how it evolved, simple observation is infeasible; there are around 100 billion stars, together with black holes, gas and dark matter. Thus, researchers turn to models. The European Space Agency has launched a craft called Gaia to measure the position and motion of a billion stars – which still represents only one percent of the total. The first data are due in 2017 and these will be used to generate models that will aim to 'fill in the gaps' and construct a map of the galaxy. From this, a model of the galaxy's evolution can be developed.

Emerging Trends and Synergies

Participants at the workshop discussed emerging trends, the way forward in computer modelling and possible areas of synergy.

1. Machine Learning

There is an increasing move towards using knowledge-based techniques, in particular machine learning, in the development of models, where algorithms are constructed that can learn from data rather than follow 'first principles' that describe the fundamental laws of nature. Many modellers are turning to machine learning as the next step in the development of models, so there is a need for training in this area. One way of obtaining greater involvement of people with expertise in machine learning may be to launch 'crowd-sourced' competitions and challenges, where models can be tested against a 'blind' benchmark. In addition, it may be advantageous to demonstrate the utility of different types of predictive approaches and algorithms in various application areas.

2. Multi-scale Modelling

Models can also be made flexible to include a wide variety of different types of data. There are clear emerging trends in multi-scale modelling; a weakness is the lack of standardisation. There are also many potential synergies between disciplines, where sampling schemes, optimisation strategies and data visualisation may be transferable. Large, multi-scale systems, such as Earth system modelling and genomics, may be able to learn from each other. Important synergies may be achieved through the blending of disciplines, notably with Earth system modelling and social and economic modelling.

Many phenomena which are modelled involve modular systems and the interaction of components into larger networks. This means that the representation of networked systems is an important and emerging area, as is the need for optimisation of models in a given parameter space.

3. Reproducibility of Research Results

New approaches are emerging in the area of data quality. Data that form the basis of knowledge-based models need to be of highest quality; therefore, it is crucial that data remains open to re-analysis and review to test their robustness. It is insufficient to rely simply on published data as these may contain biases, and, for this reason, it is important that complete raw datasets are made available for analysis, as well as the scripts that were used to analyse the data: in other words, the complete 'computational pipeline'. Because data analysis usually entails multiple steps, it can be difficult to re-construct these in a meaningful way. Thus, the best way forward is that datasets, scripts, and computer programs are not only deposited, but also formatted and documented in a form that can be used by other researchers. In this context, sharing raw datasets and scripts is expected to result in an increased reproducibility of scientific results. By using these practices, the modelling community can set standards for data quality, benchmarking and reproducibility. In order to do so, the establishment of continuous and extensive links between those collecting the experimental data and those developing new modelling approaches will facilitate data sharing and aid overall data quality.

Initiatives that aim to produce greater reproducibility in research are already running. The Research Data Alliance[1] (RDA), for example, is an international initiative whose vision is for researchers and innovators to openly share data across technologies, disciplines, and countries to address the grand societal challenges. Another effort, Research Data Canada,[2] is an example of an initiative that has been established to train scientists from an early stage in data stewardship, including in how to manage data in a way that ensures they are compatible with the concept of sharing. The Synapse[3] project in Seattle is a platform to support open, collaborative data analysis, with the aim of producing clear, reproducible science. It currently targets scientists working with clinical and genomics data. Similar efforts[4] are being undertaken by various universities in the UK.

An important issue is that of uncertainty, since databases and models contain inherent uncertainties. It is crucial that these are made explicit when data and models are deposited. As such, a caveat contained within the citation that describes the uncertainties plays a key role in of the description of each data record. Moreover, the emerging trend is to publish models that include information about the model's precision and accuracy, and the distinction between precision and accuracy should be made clear.

Forecasting and prediction are emerging trends for modellers and possible areas for synergy with other disciplines. In this context, it is important to have the input of mathematicians, statisticians and computer scientists as well as social scientists, humanities experts and economists. Often the uncertainties of models for these uses are not widely appreciated and there should be clear caveats. It may be possible to develop a standardised way to quantify uncertainty in predictive models.

Standardisation of data generated from experiments needs to be revisited. For example, samples collected at one location are extremely unlikely to be the same as those collected at another location, given the many sources of potential error in varying protocols, sample collection and storage methods, and so forth. This means that experimental data are usually not transferable and cannot be combined from different sources. This requires a major effort and, where it is not feasible or possible to achieve absolute standardisation, there must be at least an increased awareness of the issue.

4. Careers and Training

New approaches to careers and training have yet to be developed for the community of computer modellers. Scientists need to be appropriately versed in issues of statistics, data analysis and modelling. The development of a useful model requires considerable support from software experts, statisticians, data analysts and managers, among others. There is a need to develop a career path for people engaged in these activities, and there is a demand for trained personnel. In the astrophysics community, for example, there have been moves to recruit people to give support in areas such as software and algorithm development. It remains difficult to find experts who can bridge the divide between data and science. One suggestion is to introduce a dedicated qualification, such as a Master's degree, in data science and/or modelling. A dedicated MSc in modelling could encompass training in diverse techniques for data analysis and simulation that are used in different fields of application and at different scales. Techniques such as analytical modelling, stochastic modelling, the use of coarse-grained and multi-scale modelling, and methods for statistical analysis of big data are of crucial importance for such a training scheme.

Moreover, statisticians and modelling scientists need to interact with the experimental scientists who collect the data, in order to understand the nature and the properties of the data and the type, the level of complexity, and any open questions related to the problem that is being studied. Data scientists may acquire valuable additional experience whilst being exposed to the 'wet laboratory', and experimentalists should be able to gain expertise in computational data analysis.

In certain fields of science, such as ecology or systems biology, it is difficult to publish a model as an integral part of original research. Models are often 'relegated' to the methods section or supplementary information as journals are interested only in the application of the model. Many papers are published

that describe results derived from models, but with little or no information about the actual model that was used to obtain the published results. This is a situation that needs to be improved in order to give full visibility to the computer modellers and to demonstrate their career achievements.

In this context, the Scientific Committee for Life, Environmental and Geo Sciences has already published an Opinion Paper entitled, 'Career Paths in Multidisciplinary Research'[5]. This paper contains an enhanced set of indicators for evaluation of scientific output. Besides publications and patents, it suggests assessing scientists based on their contribution to the development of freely-available data access models, repositories, webometrics, and enabling tools such as methods, algorithms and software.

Recommendations

Workshop participants formulated a number of recommendations for future actions, with the hope that the suggestions in this report will be a source of inspiration for the research community as well as for research policy organisations.

It was agreed that computer modelling should become an integral part of the research process, and that funding schemes should be targeted at producing better data, dedicated people, clearly-defined research goals, better statistics and reproducibility. This, in turn, will lead to improved insights into science. In addition, it is important to include many fields of expertise in the development of computer models, for example humanities and social sciences in environmental science, and maths, engineering, computing and physics in biological sciences. To demonstrate to policy makers the importance of modelling, the community of computer modellers should gather a portfolio of studies which show how modelling has been central to the development of an outcome that has benefited society and/or science. However, decision makers must be aware that computer modelling will always remain a simulation of reality.

Specific recommendations are:

Data, Methods and Research

- Experimental scientists should consult closely with statisticians when designing experiments, especially for large-scale data collection such as in medical studies;
- Statisticians and data modellers should work in a team with experimentalists in order to receive first-hand feedback and an appropriate description of the nature and features of the experimental data and of the scientific questions the experimentalists are seeking to address;
- Scientists should share raw data in repositories and should make the software available, in a useable form, to allow re-analysis and the opportunity to test the research results. Crowd-sharing may be a useful way to promote reproducibility in science; and
- Where a model is published it should contain information on uncertainty, precision and accuracy.

Research Infrastructure

- There is a need for dedicated modelling infrastructures combined with continued support for data acquisition; and
- Research infrastructures in systems biology need further support.

Careers and Training

- Credit and recognition should be given to researchers involved in creating databases, repositories and relevant software;

- There is a need for more training in the following areas:

  - Machine learning across different disciplines using different types of model;
  - Representation of networks and the development of ontologies;
  - Methods for optimisation of models; and
  - Statistical methodology and data analysis.

- The possibility of a dedicated Master's degree in modelling should be considered. Such an MSc should encompass training in diverse techniques for data analysis and simulation that are used in different fields of application and at different scales. Techniques covered should include analytical modelling, stochastic modelling, the use of coarse-grained and multi-scale modelling, and methods for statistical analysis of big data.

Funding

- Resources should be made available for the creation and maintenance of databases/ repositories; and

- Resources should be available for the development and maintenance of analytical software; and

- Research should be done into how best to create self-sustaining resources, for example looking at ways to build a community of researchers around a software tool where many people contribute.

Possible Follow-up Activities

The creation of a portfolio of case studies showing the importance of modelling to science and society would be potentially valuable. Examples might include epidemiology for prediction and prevention of disease, the use of models in drug development, and many others.

Emerging future potential topics include:

- Reproducibility in research – including a consideration of the issues of making raw data available, software sharing and best practice;

- Model optimisation and methodologies for making a 'good model'; and

- Models for education and training.

References

[1] https://rd-alliance.org/
[2] http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca/eng/
[3] http://sagebase.org/synapse-overview/
[4] http://researchdata.ox.ac.uk/
[5] http://www.scienceeurope.org/uploads/SCsDocuments/LEGS_Careers_OpinionPaper_FIN.pdf

Annex

10 December 2014, Hotel Metropole, Brussels

09:00-09:15 Welcome and Introduction to Science Europe, the work of the Scientific Committee, and the Workshop: Dr Magdalena Radwanska, Science Europe Scientific Committee for Life, Environmental and Geo Sciences

09:15-09:30 Aims and Expected Outputs: Professor Janusz Bujnicki, International Institute of Molecular and Cell Biology (IIMCB), Warsaw, Poland

09:30-09:45 'Tour de Table'

09:45-14:30 Case Studies Demonstrating how Modelling is Providing New Insights into Science

- **Systems Biology of Cancer:** Professor Gary Bader, Donnelly Centre, University of Toronto, Canada
- **Global Nutrient Cycles and the Earth System:** Professor Benjamin Houlton, University of California, Davis, USA
- **Modelling Galaxies:** Professor John Magorrian, University of Oxford, UK
- Questions and Discussion

Moderators: Professor Janusz Bujnicki, IIMCB, Warsaw, Poland and Dr Magdalena Radwanska, Science Europe

- **Temple-based and Temple-free Modelling of RNA 3D Structures - Inspired by Techniques for Protein 3D Structure Prediction:** Professor Janusz Bujnicki, IIMCB, Warsaw, Poland
- **Multi-scale Approaches for the Simulation of Biomolecular Systems:** Dr Valentina Tozzini, Institute of Nanoscience, Pisa, Italy
- **How to Model Chemical Kinetics: Bottom-up (Inductive) or Top-down Deductive:** Professor Guy B Marin, University of Gent, Belgium
- Questions and Discussion

Moderators: Professor Lucia Banci, the Centre of Magnetic Resonance (CERM), University of Florence, Italy and Dr Vincent Reillon, Science Europe

- **Systems Immunology and Modelling:** Professor Yvan Saeys, University of Gent, Belgium
- **Molecular-level Modelling of Biological Processes:** Professor Lucia Banci, CERM, University of Florence, Italy
- **Forest Ecosystems:** Dr Gaby Deckmyn, University of Antwerp, Belgium
- **Plant Systems Biology:** Professor Steven Maere, University of Gent, Belgium
- Questions and Discussion

Moderators: Professor Ruedi Aebersold, University of Zurich, Switzerland and Professor Hojka Kraigher, Slovenian Forestry Institute (SFI), Slovenia

15:00-16:30 General discussion

- **Challenges and Opportunities: Type and Data Quality, Analysis, Integration, Methodology, Infrastructures, Careers and Training**
- **Emerging Trends**
- **Synergies and Co-operation**

Moderators: Professor Janusz Bujnicki, IIMCB, Warsaw, Poland

16:30-17:00 Recommendations

Moderators: Dr Magdalena Radwanska, Science Europe

17:00-17:15 Closing remarks: Dr Bonnie Wolff-Boenisch, Science Europe

Organising Committee: Janusz Bujnicki, Lucia Banci, Ruedi Aebersold, Hojka Kraigher, Bonnie Wolff-Boenisch, Magdalena Radwanska

Science Europe is a non-profit organisation based in Brussels representing major Research Funding and Research Performing Organisations across Europe.

More information on its mission and activities is provided at: www.scienceeurope.org.

To contact Science Europe, email office@scienceeurope.org.

SCIENCE EUROPE
Shaping the future of research

Science Europe
Rue de la Science 14
1040 Brussels
Belgium

Tel +32 (0)2 226 03 00
Fax +32 (0)2 226 03 01
office@scienceeurope.org
www.scienceeurope.org