The implications for Bioinformatics of integration across physical scales

T.C. Hodgman[1], Y. Ugartechea-Chirino[2], G. Tansley[3] and I. Dryden[4]

[1]Multidisciplinary Centre for Integrative Biology, School of Biosciences, University of Nottingham, UK, LE12 5RD
[2]School of Biology, University of York, UK, YO10 5YW
[3]School of Mechanical, Materials and Manufacturing Engineering, University of Nottingham, UK, NG7 2RD
[4]School of Mathematical Sciences, University of Nottingham, UK, NG7 2RD

Abstract

Bioinformatics blossomed with research developments in molecular biology. But as the focus of research moves back up the physical scale to the biology of whole multicellular organisms, there are new integration challenges. Data integration is a perennial theme that will not be explored here. Rather, we outline a survey of the types of data generated in whole organisms, putting particular emphasis on complex image capture, management and analysis, bio-systems modelling techniques at this higher scale, and approaches to integrating models across these physical scales. This change in biological focus raises certain challenges, one of which may be the need to retrain bioinformaticians in sciences pertaining to whole plants and animals.

1 Introduction

Most people consider that Bioinformatics developed when computers were powerful enough to assist in the storage and analysis of the sequences and structures of individual genes, RNAs and proteins. With improvements in laboratory technologies, the generation and throughput of such data required novel computational approaches to keep pace with these developments and the ability to work with genome sequences in their entirety. The next development concerned post-genomic technologies (i.e. transcriptomics, proteomics, and metabolomics) which investigate large populations of biological molecules. Indeed, the "-ome" suffix implies the levels of *every* biological molecule within the sample. It is also no surprise that people have developed computer models of the dynamics of biomolecular systems.

The latter face several challenges arising from a lack of knowledge. Inference methods can assist in identifying potential molecular interactions that have not been previously reported – identifying novel edges between network nodes. However, parameter values relating to the numbers (or concentrations) of molecules and kinetics of molecular interactions are far harder to determine. One approach can be to apply constraints, but this leads to the question of what are reasonable values. Taking data from a higher physical scale can provide upper limits (14)

Furthermore, laboratory technologies are still moving on. Microscopy, magnetic resonance and tomographic images of cells, organs and entire organisms are now being stored electronically; and this new (complex image) datatype is the subject of research by a growing number of computer scientists. This higher physical scale is an area in which Bioinformatics research is likely to become increasingly active, given its focus on applying (new) informatics to (new) biological data. This also takes dynamic modelling to a different scale, in which physiology, pathology and environmental interactions are the primary biological subjects. However, it is clear that integration with molecular-scale processes will be most incisive as

that is the level at which chemistry and mechanical interactions are taking place. In one direction, the molecular scale indicates which cells are capable of doing what, while the cytological and whole-organ data provide constraints for molecular-scale models.

This article introduces the various technologies and computational techniques, the leading multi-scale modelling projects, the informatics issues that are likely to arise and hence the challenges ahead. There will also be passing reference to our first steps in developing models of a virtual root of *Arabidopsis thaliana*.

2 Higher-scale Technologies

Various forms of imaging generate most data on higher scale structures and behaviour, though there are other physical techniques employed as well. From the naked eye, image data are being generated along three lines of development: microscopy in various forms, scans or spectroscopy in different physical ways (e.g. MRI magnetic resonance imaging, PET positron emission tomography, or CT computer tomography), and aerial or satellite imaging to detect environmental features. Most of this article will refer to the first of these and many of the computational issues also apply to the second. The third is outside the scope of this work. Microscopy provides the link between what is happening at the molecular level with scales that concern most people. At progressively higher magnifications, we can resolve the organisation of cells into tissues, cell compartments, and organelle structures and distributions. Some of the latter include multi-molecular complexes. The biggest technical developments in recent years to increase the resolution and value of these data are confocal microscopy, on which many books have now been written (e.g. ref 10), and the development of a range of specific tags that fluoresce at different wavelengths.

Confocal microscopy resolves finer structure details because it collects light from only a thin section through the total sample. Light from other sections does not obscure the detail. It is, therefore, also possible to take a series of "optical" sections and reconstruct within the computer a 3D representation of the sample (5). A fourth dimension can be introduced by capturing images at regular time intervals. This essentially generates a video that shows changes in shape, composition, size or position over time. Higher dimensions can then be added through the use of fluorescent tags. The latter include fusion proteins with green or yellow fluorescent protein, biotin-linked proteins or polynucleotides to which labelled avidin binds, or dyes that bind to membranes, carbohydrate or DNA (see figure 1). Such images generate a large amount of data on the distribution and intensity of a limited number of molecular entities, and the technique is increasingly known as high-content, as opposed to high-throughput, analysis.

Similar multidimensional data are being generated in MRI, CT and PET scans. The first two of these show anatomical changes or movements over time, while the latter is more relevant to metabolic studies, e.g. it is now used routinely to look at oxygen utilisation and glucose metabolism in areas of the brain as an early indicator of predisposition to Alzheimer's disease (2).

There is also a broad range of other physiological and mechanical data that can be determined and used for modelling. In animals these include such things as ECG traces, blood pressure, oxygen utilisation and a broad sweep of bio-assays. From plants, we can determine such things as transpiration rates, xylem and phloem exudation rates and turgor pressures within cells. Although many of these techniques do not generate large quantities of data, they do provide indicators of an organism's state, which will have a bearing on molecular-scale processes, e.g. the level of exercise or stress may be high enough for a cell to switch from aerobic to anaerobic metabolism.

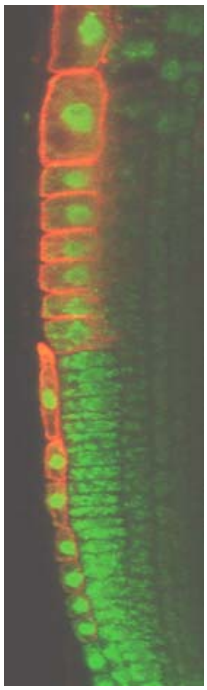

Figure 1. This is a laser confocal microscope section through part of an *Arabidopsis thaliana* root tip. The green and red dyes respectively show the locations of vacuoles and a cell-surface hormone transporter.

3 Image analysis techniques

Once these images have been generated, there is then a range of steps to be carried out to extract relevant information (see ref. 12 for detailed fundamentals). Each image comprises and array of pixels which carry information about the level of each colour (red, blue, green) and brightness. From the raw image, various steps can be carried out to reduce the file size including reducing the total number of pixels through amalgamation of pixel values by various criteria, or reducing the number of colour or brightness levels. This is *image compression*. *Image enhancement* involves a range of techniques including adjusting colour/brightness levels to reveal greater texture/detail, applying maximum entropy techniques to improve clarity or applying algorithms with assumptions on motion-induced blurring to sharpen edges.

From this point, the are well established algorithms to look through these matrices for edges between boundaries of colour/brightness, which leads on to *segmentation*: the separation of the image into a series of zones whose features (shape, colour, granularity) can be individually determined. More advanced computer vision seeks to assign meaning to individual segments on the basis of some reference data (e.g. cell shape and colour can suggest what type of cell it is). This really comes into its own when image warping used to fit one image onto an annotated reference, as might best be exemplified by functionality in the Edinburgh mouse atlas project (EMAP), outlined further below.

Libraries of images can also be used as a dataset for statistical shape analysis (ref. 3, see also figure 2). A range of reference points on a common shape can be subject to analyses to find out which distributions of positions are statistically meaningful and hence which can be used to classify differences between related shapes.

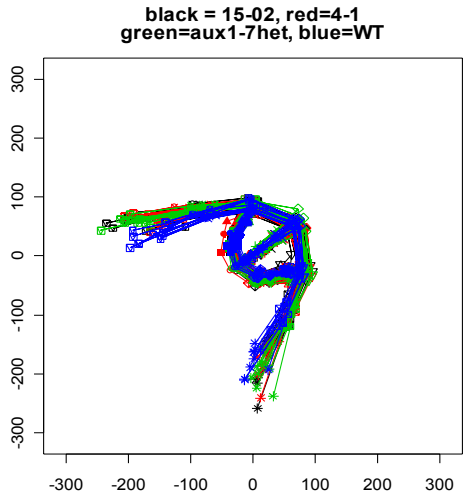

Figure 2. Procrustes analysis of *Arabidopsis* root tips in wild-type (blue) and three mutant strains (black, red and green). In the case of all 3 mutants, the root tips up to the lateral root caps are blunter.

4 Modelling techniques

Since morphology is more important and diffusion a much more significant factor in processes at the cellular scale and above, it is not surprising that the mathematical and computational techniques used for molecular-scale systems-modelling are less appropriate. Bioengineers have used well established techniques, such as Finite Element Analysis, to define cells or zones of tissue and their interactions. This is best exemplified by the Physiome Project (6). Boundary Element analysis is a similar modelling technique but pays specific attention to the interfaces between elements (1), and several computational physiology groups have applied Computational Fluid Dynamics to simulate biological systems. The final modelling technique to be applied in this context involves cellular automata (8,13). People may be familiar with this technique through "Game-of-Life" screensavers, or modelling packages like Stochsim (9). These involve cells in an array changing their properties in defined ways and in response to the states of neighbouring cells. They have been used to model both molecular systems and developmental processes such as micro-vascular growth. (15).

5 Integration across the physical scales

There is a growing number of exemplar projects involving multi-scale modelling. The most developed is the virtual heart (11). It involves a subset of Ordinary Differential Equations defining the energy status, ion channel behaviour and muscular contraction signalling of individual cells. These are embedded into one FEA model (containing >1000 elements) to show the waves of depolarisation as they move through the heart tissue. This model generates validated ECG traces when electrical potentials across the virtual heart are calculated and has been used to support an application to the Federal Drugs Administration for licensing a new medicine. It also predicts the effects of inappropriate biopsies in causing the heart to flutter and the effects of minor changes in ion channel activity on overall behaviour of the heart. A different FEA model splits the heart into ~50 elements, in which the orientation of muscle fibres has been experimentally determined. These data enable every element to warp in response to muscular contraction, resulting in a validated pumping-heart model. A range of

other human tissues are being subjected to the same multi-scale development, including lung, digestive tract, kidneys, skeleton, and liver.

The Edinburgh Mouse Atlas Project (EMAP) is an integrative biology environment in its own right. Serial sections of mice embryos have been taken and subjected to microscopy to create a 3D image dataset – each element a known as a voxel. An anatomist has worked through the 3D representation and assigned the appropriate anatomical terms to the different structures. When someone has taken a new serial section and subjected it to *in situ* hybridisation or immuno-localisation, the resulting image can be submitted to the EMAP software. The system first warps the "query" image to find the best plane through its 3D representation, and can therefore indicate which parts of which tissues have been highlighted by the labelling experiment. The system outputs the cells/tissues concerned for both the benefit of the laboratory scientist and for the modeller who wishes to know which gene products are present in cells of interest. EMAP is one example of how studies on a larger physical scale can feed information down to molecular scale models. In a similar vein, microbial growth rates in fermenters have provided constraints for virtual cell models, that can now be considered as validated and predictive (4).

This approach is also being used to model *Arabidopsis* growth. Work on the Computable Plant project has taken a combination of imaging and Partial Differential Equation models to uncover the interactions of a transcription factor network in the Shoot Apical Meristem (7). The virtual root programme, recently initiated by the University of Nottingham Centre for Plant Integrative Biology, is expanding on this approach. It has taken time-series and other image data to determine the growth rates of cells in the elongation zone at the *Arabidopsis* root tip. These values can then be fed into Finite Element models of the interactions between the different cell types, and provide constraints for the macromolecular synthesis models of each cell type. The cells themselves will be subject to a battery of post-genomic techniques that determine expression, protein and metabolite profiles to provide data for molecular scale models of how cells grow. These multi-scale models can implicitly simulate the growth response of specified mutants, simply by altering the parameter values for the molecular entities concerned (see figure 3). The programme will also investigate novel search and optimisation techniques that aim to infer potential genotypes and environmental conditions that result in specified phenotypes.

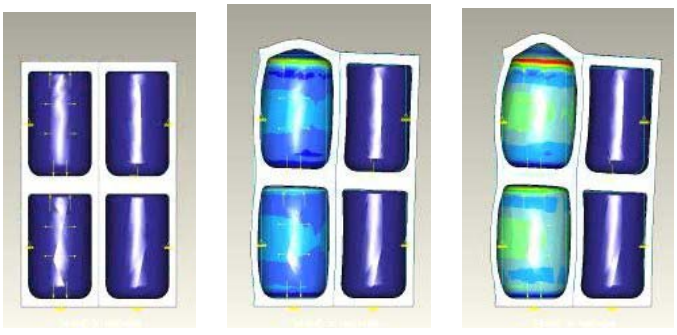

Figure 3. A four-cell model using Finite Element Analysis. The sequence depicts two cells on the left having uncontrolled expansion. Such bulging has been observed in the roots of *Arabidopsis* mutants with aberrant hormonal regulation (Bennett and Swarup, unpublished work).

6 Hardware Issues

The machines generating such organ-scale imaging data often have software directly attached to put the data into files on a local disk for further analysis. However, the principal challenge

in handling image data is file size. A single image can easily be several megabytes if its resolution size is high. A time-series or serial-section dataset may also associate hundreds of images, resulting in datasets up to 200 gigabytes (S. Mooney, personal communication). Effective management of such data requires specialist hardware (Storage Area Networks, or Network Attached Storage) and care in data transfer, as this could result in networks grinding to a halt or image analysis work progressing unacceptably slowly.

7 Informatics issues

Bioinformatics developments for cell and organ scale data have been uneven, probably because much of the work is being carried out by a small number of groups. The primary source data are often images and the issues relating to high data volumes have already been discussed. A worldwide-web search reveals that there are now many "image databases" accessible online, but they are very largely repositories with little supporting metadata or connectivity to other information resources. Each image/dataset should have associated data on the magnification and imaging settings, the conditions under which the biological sample was prepared (which itself ought to conform to a defined Standard Operating Procedure), and management details, such as the date, time and name of the person producing the data. These extra data should conform to a standard ontology or controlled vocabulary so that the most sophisticated analyses can be carried out. At this time, we are not aware of any such standards. Regarding analysis of image data, the current software is largely limited to systems that accompany the hardware, or bespoke implementations developed by computer-vision scientists. There is much work to be done to make a versatile image-analysis environment that can be used intelligently by life scientists.

The models themselves must conform to data standards if they are to be useable in proper informatics environments. Although Systems Biology Markup Language might suffice for some molecular-scale models, it will not serve so well for capturing higher-scale physical details, such as turgor pressures or sheer stress. Members of the Physiome consortium have developed ontologies for tissues, anatomy and physiology (see Hunter and Borg 2003 for further details), but they have a mammalian emphasis, so new or extensions to current standards may be required.

8 The challenges ahead

The above discussion has outlined how data from higher physical scales can be used for inferring the values of parameters in molecular-scale models. However, the latter can indicate maximal growth rates that could be used in higher-scale (e.g. Finite Element) models to reveal how organs grow and acquire their overall morphology. Integration of these approaches would involve processes (on the same or different CPUs/servers) communicating relevant data to each other. This is one potential approach to multi-scale modelling, which is currently one of the biggest challenges in systems biology development. It remains possible that other approaches may be the way forward. One such example is hierarchical models, in which molecular-scale model simulations (say of virtual cells) running on one or more CPUs communicate or receive parameters from simulations on other CPUs which take cellular or whole organ structure into account. Software agents might be used to achieve this effect. However, it is certain that the data must be in a suitably structured format if it is to conform to good informatics standards and be readily useable in different modelling frameworks.

The second major challenge lies in the skill-sets of bioinformaticians themselves. The biological focus of most bioinformatics development has been the capture, management and analysis of molecular data. The move to integrative biology will require informaticians who

are also familiar with biological systems at higher physical scales: anatomy, physiology, pathology, population biology, ecology, and perhaps even psychology if brain behaviour is the subject of investigation. Hence, we can expect a need to retrain our bioinformaticians in these new areas of biology so that they can apply their current skills to greater impact in the future.

9 References

1. Beer, G. (2001) *Programming the boundary element method*. John Wiley & Sons.
2. Brooks, D.J., Lammertsma, A.A., Beaney. R.P., Leenders, K.L., Buckingham, P.D., Marshall, J., Jones, T. (1984) Measurement of regional cerebral pH in human subjects using continuous inhalation of 11CO2 and positron emission tomography. *J Cereb Blood Flow Metab*. **4**, 458-465.
3. Dryden, I.L. and Mardia, K.V. (1998) Statistical shape analysis. Wiley& Sons, Chichester, UK.
4. Edwards, J.S., Ibarra, R.U. and Palsson, B.O. (2001) *In silico* prediction of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature Biotech.* **19**, 125-130.
5. Haseloff, J. (2003) Old botanical techniques for new microscopes. *Biotechniques* **34**, 1174-1182.
6. Hunter, P.J. and Borg, T.K. (2003) Integration from proteins to organs: the physiome project. *Nature Reviews Molecular Cell Biol.* **4**, 237-243.
7. Jőnsson, H., Heisler, M., Reddy, G.V., Agrawal, V., Gor, V., Shapiro, B.E., Mjolsness, E. and Meyerowitz, E.M. (2005) Modelling the organization of the WUSCHEL expression domain in the hoot apical meristem. *Bioinformatics* **21 S1**, i232-i240
8. Kansal, A.R., Torquato, S., Harsh IV, G.R., Chiocca, E.A. and Deisboeck, T.S. (2000) Simulated brain tumor growth dynamics using a three-dimensional cellular automaton. *J. theoret. Biol.* **203**, 367-382.
9. Le Novère, N. and Shimizu, T.S. (2001) Stochsim: modelling of biochemical stochastic processes. *Bioinformatics* **17**, 575-576.
10. Muller, M.A. (2005) Introduction to Confocal Fluorescence Microscopy. Society of Photo-Optical Instrumentation Engineering
11. Noble, D. (2002) Modelling the heart – from genes to cells to the whole organ. *Science*, **295**, 1678-1682.
12. Peterou, M and Bosdogianni, P. (1999) Image Processing – the fundamentals. John Wiley and Sons, Chichester, UK
13. Sipper, M. (1999) The emergence of cellular computing. *Computer* **32**, 18-26.
14. Von Lieres, E., Petersen, S., and Wiechert, W. (2004) A multi-scale modelling concept and computational tools for the integrative analysis of stationary metabolic data. *J. Integrative Bioinformatics* 4.
15. Yingling, M., O'Neill, T., Skalak, T.C. and Peirce-Cottler, S. (2005) *Proc. 2005 Systems and Information Engineering Design Symposium*.