Molecular Systems Biology (2006) doi:10.1038/msb4100080

www.molecularsystemsbiology.com
Article number: 40

molecular systems biology

EDITORIAL

Dry work in a wet world: computation in systems biology

Molecular Systems Biology 4 July 2006; doi:10.1038/msb4100080

Systems biology: a new term for an old science

Prior to the outburst of molecular genetics in the latter part of the past century, studying biological systems in their whole was commonplace owing to the limited scientific knowledge and appropriate molecular tools available. The molecular understanding of each observable phenotype was uncertain and based on empirical deductions from complete systems (Von Bertalanffy, 1950; Kacser and Burns, 1973). Large-scale molecular biology has led to the routine deciphering of genome sequences and the subsequent identification of gene products and metabolites. Using these molecular reagents, thousands of studies have drawn novel molecular mechanisms, defined signalling cascades and molecular interactions. However, despite the ever-increasing ability to catalogue the players in biological systems, the relationship between overall behaviour of the biological system and the newly discovered molecular mechanisms remains often puzzling and elusive. This is relevant not only to the general understanding of biology but also to its application in understanding human disease. The production of a 'disease' phenotype is the result of many interacting components, from 'simple' monogenic diseases through to complex disease with multiple genetic and environmental factors. If we wish to define the molecular targets appropriate for therapeutic intervention, develop better diagnostics and understand how environmental factors influence disease, we must return to the study of complete biological systems armed with the ever expanding catalogue of molecular reagents. Systems biology is a new term for the old science of studying biological systems holistically, but now reinforced with high-throughput, increasingly affordable molecular tests and considerable sophistication in computational modelling (Kitano, 2002). We will focus here on the informatics of systems biology and how this rediscovered discipline is changing the way computers are used in molecular biology.

Where computational systems biology meets experimental systems biology

Organized merging of computational systems biology with experimental systems biology is the most important challenge facing modern laboratories. We categorize the informatics into three distinct layers, each of which needs its own expertise and investment, but fundamentally provide information into the next one. The layers are (i) close-to-data generation informatics, (ii) large data sets informatics and (iii) systems modelling informatics.

Close-to-data generation informatics

Experiments that inherently handle many data points require their own software and databases specific to the experimental system. Most equipment (e.g. microarray platforms, proteomics, automated microscopes, among others) come with their own computer and software bundled with the instrument. The standardization of these computational tools is very variable depending on the maturity of the overall industry and the investment by both the instrumentation company and early adopter sites. Any laboratory using such equipment will need increasing sophistication in computational methods with personnel who are competent in extracting, moving and troubleshooting data sets in a computational setting. We estimate that between 20 and 30% of the salary resource should be dedicated to close-to-data production informatics. These include mostly database generation and maintenance tools such as Laboratory Information Management Systems (LIMS). Examples of data management system are SBEAMS (Institute for Systems Biology) or BASE, for microarray analysis (Saal *et al*, 2002).

Bioinformatics of large data sets

Modern high-throughput technologies produce large amounts of data quickly, and their storage and organization require informatics resources. Databases can be separated into two overlapping categories, that is, experimental databases such as Arrayexpress for microarray data (Parkinson *et al*, 2005) or IntAct for protein–protein interactions (Hermjakob *et al*, 2004) on the one hand, and knowledge databases on the other hand, to which the first ones can be compared, interpreted and integrated. Such knowledgebase can consist, among others, of attributes of proteins mined in-depth from the literature such as Uniprot (Bairoch *et al*, 2005), or qualitative molecular reactions and signalling pathways such as Reactome (Joshi-Tope *et al*, 2005). These data sets are most useful when made public, allowing other groups to query and integrate their data. To do so, there is increasingly sophisticated bioinformatics software such as Cytoscape (Shannon *et al*, 2003) and Bioconductor (Gentleman *et al*, 2004). Despite the availability of these data sets and software, we believe that some critical components in the integration of public data sets with 'local' private data are currently underdeveloped. This is part of the focus of our project, ENFIN, described below.

Systems biology modelling

One of the most effective ways to investigate the properties of a given system is to build a computational model of it. The computational model can then be used to simulate system's behaviour and potentially reveal whether experimental results are consistent or not with the model. A researcher can build a deeper understanding of the system by iterating between model design, testing and validation using experimental challenges. This is particularly true when an emergent behaviour of the model cannot be easily explained by simple 'presence/absence' arguments, such as a meta-stable switch-like response between two states triggered by a threshold of an input signal. The development, theory and use of these computational models represent the third computational area in systems biology (Figure 1).

Quantitative and qualitative modelling of biological systems started with the ones of metabolic processes in cells, which remain a strength in the field. These models can be based on stoichiometric properties of metabolic pathways (Kacser and Burns, 1973) or, in the case where more detailed parameters are available, kinetic modelling as a series of differential equations (Le Novere *et al*, 2005; Kowald *et al*, 2006). These models become quite sophisticated and there is a growing endeavour of cataloguing, testing and aggregating them, for example using Systems Biology Markup Language (SBML) and databases of models, for example, BioModels Database (Le Novere *et al*, 2006) or CellML (Lloyd *et al*, 2004). Kinetic modelling is not the only framework aiming at understanding biological systems. Simpler models based on Boolean logic gates or simpler analogies with electrical circuits allow more complex systems with fewer known parameters to be tackled (Thomas, 1973; Savinell and Palsson, 1992; Sanchez and Thieffry, 2003). This area of modelling merges progressively with investigations of large-scale data sets, such as the use of graphical Bayesian methods, which can learn aspects of the model from large-scale data sets (Segal *et al*, 2005).

Although there is considerable sophistication in the construction and use of these modelling methods in specific computational groups, they remain mostly forsaken by experimentally focused biologists. This is mainly owing to engineering aspects of how to interact and construct these models, as well as exchange of knowledge between the computational groups and experimental groups, both potential providers of useful modelling frameworks on the one hand and of useful experimental data for these frameworks on the other hand. Again we hope to address some of these shortcomings in ENFIN.

Integration across disciplines

Whereas close-to-data production informatics demands skills in data reformatting, large data set handling and systems simulation require analysis bioinformatics and computational modelling aspects. In practice, these layers may be integrated via collaborations, a local bioinformatics or systems biology group, or by people with mixed skills in experimentation and informatics, including abilities to script, to use sophisticated statistical tools (R, S, among others) and to understand where the appropriate biological data sets are.

We have recently started a new project, ENFIN, a European Network of Excellence to address some of these challenges. ENFIN (www.enfin.org) is mainly a bioinformatics group, with multiple disciplines distributed across mathematics, computer sciences and biology. We are mainly focusing our efforts in the latter two layers of computational systems biology, namely the integration of public data with local experimental data and systems biology modelling. In addition to 15 computational groups, ENFIN comprises five experimental groups, ranging from investigations into bacterial metabolism through to mammalian cell signalling. These experimental groups will be providing both interesting problems and motivation for the computational groups, which

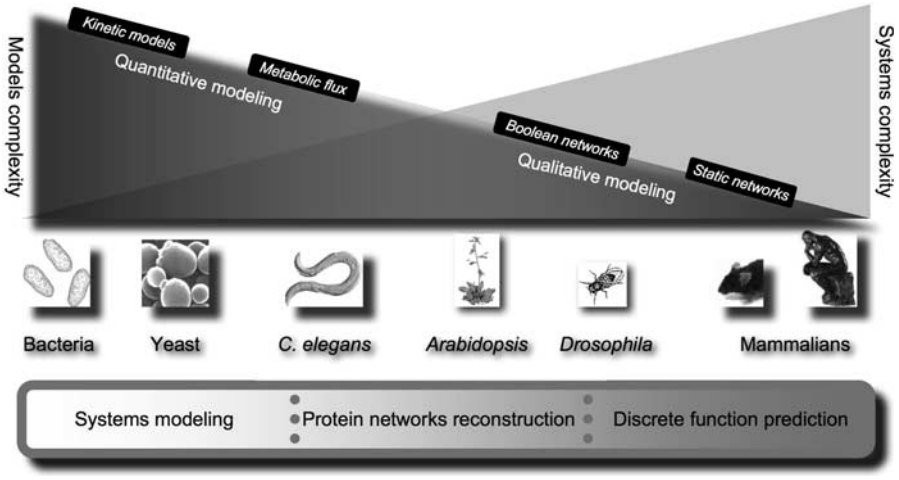

Figure 1 Schematic representation of the methods currently used to model biological systems of various complexities. Many of the more sophisticated models require numerous kinetic parameters (or constraints of these parameters) and accurate measurement of the concentration of many of the biochemical species. One could imagine such models being effectively combined with experiments on relatively 'simple' systems such as bacteria and single-celled eukaryotes. In contrast, it is far harder to imagine these detailed models being effective for multicellular eukaryotes, which both have the challenge of different cell types and multiple intracellular compartments. Qualitative models seem rather more appropriate for metazoan organisms. These include Boolean models and graphical Bayesian models. In both cases, simpler and noisier data sets can be used to inform or train the models.

will in turn explore the possibilities of expanding the empirical knowledge by applying methods to predict protein function, reconstruct protein networks and modelling systems; consistent with our view that close-to-laboratory informatics is crucial for systems biology, each experimental group is hiring a dedicated bioinformatician. ENFIN's infrastructure will be made publicly available and will be designed in an open manner, using Web services and other networked technologies.

The project aims at developing and integrating bioinformatics tools in a common platform to be used to analyse a high variety of experimental data sets, and to deliver to the user a panel of predictions and models virtually expanding the spectrum of a classical interpretation of biological data sets. Measuring the impact of such a project is often difficult and we will illustrate ENFIN's top-level goals using the TGF-beta pathway as an example. The TGF-beta pathway represents an interesting challenge as it produces non-trivial differences in response to various related stimuli, modulated by cellular context and other factors (Moustakas and Heldin, 2005).

In normal cells, growth factor signals (like those from TGF-beta1) are interpreted by tightly regulated networks of signal-transduction proteins controlling the appropriate cellular response. Cancer cells are often unresponsive to normal signalling cascades and show very unpredictable behaviours. The intracellular signalling machinery seems to consist of mainly two closely related Smad cascades, the TGF-beta and the BMP cascades. Various factors can influence the TGF-beta pathway: regulatory inputs towards Smads or the nucleus, such as Ras or the oncoproteins Ski and SnoN; signalling effectors downstream of Smad, such as the protein kinase A; signalling effectors downstream of the receptor, such as p38 or MAPK. Whereas the Smad pathway is rather simple and linear, complex crosstalks with other signalling modules that are also activated by TGF-beta ligands such as the PI3-kinase/Akt, the Rho GTPase and the MAPK/JNK/p38K pathways occur. The role of these alternative pathways in the diversity of cellular responses remains unclear. There is emerging evidence that acute activation of a single mitogenic oncogene in mammalian cells not only promotes cell proliferation but simultaneously turns growth-opposing cellular programmes on. These include apoptosis and an irreversible growth arrest termed premature senescence (Kahlem *et al*, 2004). Finding either genetic or pharmacologic means to tilt this balance towards cellular senescence or apoptosis would provide an intriguing opportunity for clinicians to selectively target oncogene-expressing cancer cells to destruction.

The use of ENFIN as an integrative platform of public data with known pathways will provide valuable material for identifying new potential pathways components. Parameters such as post-translational modifications or enzyme:substrate putative interactions will be investigated (for instance, producing a ranked list of possible phosphorylation processes involving a given kinase). The considerable knowledge collected on the TGF-beta pathway is not yet available in a computational form; ENFIN will facilitate its description, allowing other tools to automatically analyse large-scale experimental data sets. Finally, it is likely that ENFIN will use a Boolean model of the signalling cascade in the first instance to create a predictive model of the pathway. This model will probably be deficient in some areas and may suggest either specific experiments to resolve particular aspects or impute specific missing components for the model to be stable. These hypotheses will be presented back to the experimentalists in the ENFIN network, and new experiments, leading to a new round of integration, will be developed.

ENFIN does not just aim at understanding TGF-beta signalling, but instead uses such examples to drive both specific method development and an extensive integration across computational tools. Our goal is to make these computational approaches accessible to a broader range of experimentalists, therefore progressively growing the area of computational systems biology beyond its traditionally theoretical level, on the one hand, and introducing more 'wet' experimentalists to power these 'dry' computational tools, on the other hand.

Acknowledgements

We express our gratitude to Drs Aristidis Moustakas, Nicolas Le Novère and Wolfgang Huber for providing helpful comments. This work was supported by the European Network of Excellence contract LSHG-CT-2005-518254.

References

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res* **33:** D154–D159

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5:** R80

Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, Margalit H, Armstrong J, Bairoch A, Cesareni G, Sherman D, Apweiler R (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res* **32:** D452–D455

Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* **33:** D428–D432

Kacser H, Burns JA (1973) The control of flux. *Symp Soc Exp Biol* **27:** 65–104

Kahlem P, Dorken B, Schmitt CA (2004) Cellular senescence in cancer treatment: friend or foe? *J Clin Invest* **113:** 169–174

Kitano H (2002) Looking beyond the details: a rise in system-oriented approaches in genetics and molecular biology. *Curr Genet* **41:** 1–10

Kowald A, Lehrach H, Klipp E (2006) Alternative pathways as mechanism for the negative effects associated with overexpression of superoxide dismutase. *J Theor Biol* **238:** 828–840

Le Novere N, Bornstein B, Broicher A, Courtot M, Donizelli M, Dharuri H, Li L, Sauro H, Schilstra M, Shapiro B, Snoep JL, Hucka M (2006) BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res* **34:** D689–D691

Le Novere N, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, Crampin EJ, Halstead M, Klipp E, Mendes P, Nielsen P, Sauro H, Shapiro B, Snoep JL, Spence HD, Wanner BL (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol* **23:** 1509–1515

Lloyd CM, Halstead MD, Nielsen PF (2004) CellML: its future, present and past. *Prog Biophys Mol Biol* **85:** 433–450

Moustakas A, Heldin CH (2005) Non-Smad TGF-beta signals. *J Cell Sci* **118:** 3573–3584

Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, Farne A, Lara GG, Holloway E, Kapushesky M, Lilja P, Mukherjee G, Oezcimen A, Rayner T, Rocca-Serra P, Sharma A, Sansone S, Brazma A (2005) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* **33:** D553–D555

Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C (2002) BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biol* **3**, SOFTWARE0003

Sanchez L, Thieffry D (2003) Segmenting the fly embryo: a logical analysis of the pair-rule cross-regulatory module. *J Theor Biol* **224:** 517–537

Savinell JM, Palsson BO (1992) Optimal selection of metabolic fluxes for in vivo measurement. I. Development of mathematical methods. *J Theor Biol* **155:** 201–214

Segal E, Friedman N, Kaminski N, Regev A, Koller D (2005) From signatures to models: understanding cancer using microarrays. *Nat Genet* **37** (Suppl)**:** S38–S45

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13:** 2498–2504

Thomas R (1973) Boolean formalization of genetic control circuits. *J Theor Biol* **42:** 563–585

Von Bertalanffy L (1950) The theory of open systems in physics and biology. *Science* **111:** 23–29

Pascal Kahlem and Ewan Birney
EMBL–European Bioinformatics Institute,
Wellcome Trust Genome Campus, Hinxton,
Cambridge, UK