Published in IET Systems Biology
Received on 10th April 2007
Revised on 1st August 2007
doi: 10.1049/iet-syb:20070020

ISSN 1751-8849

Modelling biological modularity with CellML

M.T. Cooling P. Hunter E.J. Crampin

Auckland Bioengineering Institute, The University of Auckland, 70 Symonds St., Auckland, New Zealand
E-mail: m.cooling@auckland.ac.nz

Abstract: In recent years advances in the construction of mathematical models of biological systems have yielded an array of valuable constructs. The authors seek to provide a 'leading practice' method for implementing modularised kinetic mass-action models in order to obtain a number of advantages in model construction, validation and derived insights. The authors advocate the consideration of 'accounting cycles' or 'chains' to define 'functional' components and the separate consideration of 'messenger' components for mobile or diffusive molecular species. From a conceptual modularisation the authors illustrate, with an example drawn from signal transduction, a component-based formulation in the model exchange format cellular modelling markup language (CellML) 1.1 – demonstrating loose coupling between functionally-focused reusable components. Finally, the authors discuss the dilemmas associated with modelling protein-to-protein interactions, and the vision for using future CellML enhancements to resolve potential duplications when combining independently developed models.

1 Introduction

To take advantage of the ever-increasing biological detail uncovered through experimentation, mathematical modellers will benefit from access to modular model components that can be combined to leverage old knowledge when dealing with the new. The validity of this partially reductionist approach is supported by recent research that indicates that such modularity exists in vivo [1], possibly to allow subsystems to dynamically adapt to changing conditions [2]. Here, we take a recent model of a signal transduction pathway (IP3 signalling in the cardiac myocyte) and break it into distinct conceptual modules.

Modularisation of cellular models provides at least five advantages to the model developer: (1) if module boundaries are appropriately chosen, then the composite models can mirror the compositional nature of the biological system. (2) Modules representing known biological components can be isolated, and their behaviour analysed for different conditions independently of the overall model. Assessing the validity of modules against experimental results assists greatly in identifying and rectifying inconsistencies, prior to a complex and time-consuming analysis of the model as a whole. (3) Although individual modules may differ, the overall patterns employed by two unrelated modules when solving a common problem may be similar. Recent research has led to the discovery of recurring 'circuit elements' or key 'wiring patterns' in signal transduction networks [2]. Repeated solutions to common problems have been identified in diverse subcellular systems and represent 'network motifs' analogous to gene or protein sequence motifs. (4) Pathways in different cell types with different biological functions may share common modules. It may be possible to build models for different cell types by joining together modules from other pathway models and refitting parameters. (5) Multi-scale modelling such as that envisaged by the International Union of Physiological Sciences Physiome Project, which aims to link models from the nano-scale to organ, tissue or whole human models, is too complex to perform in a single effort. It is more appropriate to take advantage of modularity and construct models at particular time and space scales. These can then be treated as 'black boxes' that summarise lower level detail while linking to the level of abstraction directly above [3].

At the molecular level, there have been many valuable models of biological systems produced. Regrettably many of these models are not readily available to the community for reuse because of a lack of a sufficiently accurate description [4]. Although some researchers do make computer code or otherwise executable component-based implementations for their modules available (for good examples see [5, 6]), the

recombination of these components to form new modules often requires the editing of this code (whether textually or visually in a graphical environment) in order to ensure that molecular species are linked appropriately for the biological scenario under investigation.

Conceptually, modules are unlikely to be rigid structures, with species belonging to different modules at different times [7] (an example of this is given below). This flexibility is not easily represented in computer code formulations and can lead to conflicts when combining components from different researchers. Although some degree of 'gluing' code is likely to be necessary, ideally, existing module implementations should not have to be modified in order for previously unanticipated connections to be made. It would be more efficient if module implementations could also be relied upon to act as black boxes without the need to understand how they are coded.

We present a method of dividing modules into reusable components in a common model exchange format that will alleviate much of the need for modifications of this kind, which we illustrate by translating our modules into computer-readable cellular modelling markup language (CellML) [8] components. We discuss how component boundaries should be formed and develop a leading practice model for such implementations, to aid future model construction via the aggregation of existing components with minimal code alterations.

2 Conceptual modularisation

We will illustrate modularisation using a biological reaction schematic of an existing IP3 signalling pathway model [9] as shown in Fig. 1*a*.

We define modules as a collection of species with both relatively strong internal interactions and representing a particular biological function [7, 10]. Conceptually, we can divide the reaction schema into three functional modules, depicted in Fig. 1*b*.

The model is composed of the GPCR module (reactions R1–R6), followed by a module that describes the interactions of PLCβ, $G_\alpha GTP$ and Ca^{2+} (reactions R8–R13). These two modules are additionally linked by reaction R7, which will be discussed below further. A third conceptual module which describes the production and degradation of IP3 is shown by reactions R14–R16.

In each case, the module is defined by a set of states of the most important species for the function that the module performs, and the reactions that allow transitions between those states. The GPCR module receives an extracellular signal (the ligand) and transduces this across the plasma membrane to the inside of the cell via receptors. This module therefore represents an 'accounting cycle' for those receptors – it contains all the possible complexes of that particular species, governed by differential equations measuring their concentrations derived from the reactions between those complexes. Accounting cycles may also be mass conservation cycles of the species of interest, as can be understood by considering the differential equations for the receptor and receptor complexes R, R_l, R_g and R_{lg}, the fluxes for which sum up to zero, as shown in Table 1.

Similarly, the second conceptual module, which reads the intracellular signal and primes the key enzyme for this module (PLCβ), exhibits similar accounting properties for that enzyme via reactions R8–R12. A module does not require a cycle for an accounting relationship to hold; an 'accounting chain' (where mass is accounted for as in a cycle but a graph of the reactions in the relationship is not closed) is also possible. This is the case for the third module that describes the production and degradation of the IP3 signal which 'ends' in a sink state.

3 Component-based implementation

To take advantage of aforementioned modularisation advantages, models must adhere to standards and semantics that encourage their compatibility with one another. The model representation should also be independent of the solver algorithm and technology platform that acts on it. A CellML representation fulfils these requirements, being a human and machine-readable XML-based exchange format for mathematical models.

CellML models are partitioned into 'components' that encapsulate internal variables and mathematical relationships. Communication between components is performed via CellML 'connections', which map a given variable from one component onto another variable in a second component. Components can also reside in separate files, connected by CellML's 'import' functionality. A conceptual module may be composed of one or more CellML components. This flexible approach facilitates the combination of modules designed by different researchers, while keeping the mathematical details of the module specification encapsulated in independently constructed components [8]. Additionally, a framework for integrating models for processes at different spatial and time scales, implemented as CellML components, already exists [11].

Implementing the modules as CellML components requires forethought on how they should be formed to maximise their utility in future models. To aid reuse, we make a distinction between highly connected 'messenger' components and non-messenger functional components. Diffusible molecules, such as the extracellular ligands, calcium and IP3 have greater potential than localised species to be consumed or produced by multiple modules and, thus, should be only loosely coupled with other

doi: 10.1049/iet-syb:20070020

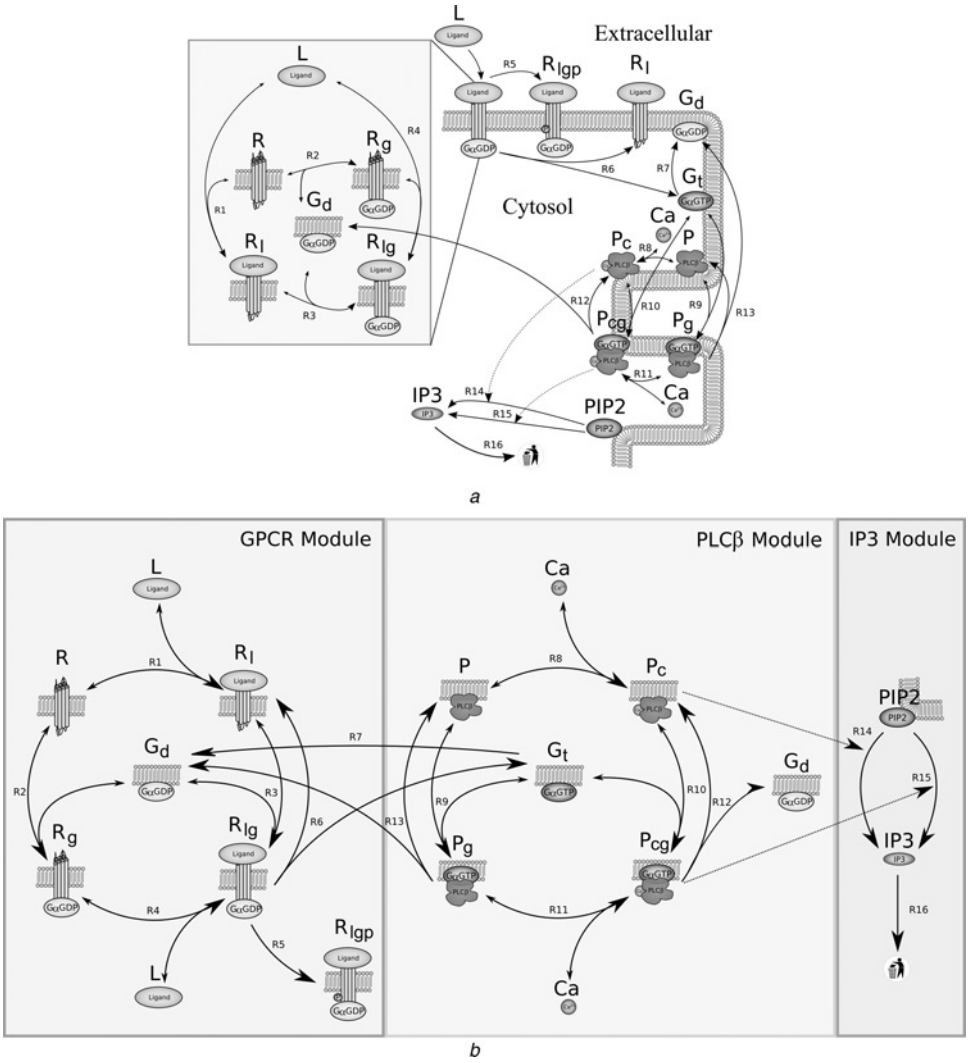

Figure 1 *Reaction scheme of the IP3 production system and conceptual model*

a Reaction scheme of the IP3 production system
b Conceptual model
Conceptually, the model is made up of three functional modules; the GPCR, the PLCβ, and the IP3 module
G_α-GDP, Ca^{2+} and Ligand species (G_d, Ca and *L* respectively) are represented twice for visual clarity
Extracellular ligand (*l*) binds to receptors (*r*), whether precoupled with G_α-GDP (G_d) or not
Fully activated receptors (R_{lg}) release G_α-GTP (G_t) which along with calcium (Ca) stimulates PLCβ (P)
In the unstimulated state, PLCβ-Ca^{2+} (P_c) hydrolyses PIP2 to produce IP3 via reaction R14
When stimulated, PLCβ-Ca^{2+}-G_α-GTP (P_{cg}) hydrolyses PIP2 at a faster rate than reaction R14 via reaction R15
Free IP3 is degraded via reaction R16
Figure 1*a* adapted from [9, Fig. 1] with permission

species. Membrane bound molecules which are nonetheless messengers, such as G_α subunits (with attendant self-GTPase reaction R7), also qualify. Hence in this formulation, all messenger molecules (ligand, G_α subunits, calcium and IP3) are given their own components.

Although it is important that components hide information that is unnecessary to other components, in the CellML framework these messenger components must expose the current concentrations of the messenger molecule species to allow these concentrations to be used in the calculation of kinetic rates inside other components. It is also necessary to allow for the connection of fluxes representing the gain or loss of messenger molecule species, because of reactions that use or produce these molecules in other components. These 'sources' and 'sinks' can be summed to a single flux $J_{\text{source,species\_name}}$ which each messenger molecule component exposes to be contributed to by other components. The summing of source and sink fluxes of other components to this single flux is implemented in an 'interface' component, which can be defined by the model-builder who connects the components together. Through these mechanisms, secondary non-messenger components can both use and

Table 1 Conservation cycle of receptors, shown by reaction fluxes (J_i, where i denotes the reaction number) for the receptor complexes that sum to zero, defines the accounting relationship for the GPCR module.
Flux directions (sign) are chosen to reflect the dominant direction under stimulated, physiological conditions [9]

dR/dt	=	$-J_1$	$-J_2$					
dR_l/d_t	=	J_1		$-J_3$			$+J_6$	
dR_g/dt	=		J_2		$-J_4$			
dR_{lg}/dt	=			J_3	$+J_4$	$-J_5$	$-J_6$	
dR_{lgp}/dt	=					J_5		
Sum:		0	+0	+0	+0	+0	+0	=0

contribute to messenger molecule concentrations defined in separate messenger components, without requiring any changes to their own internal CellML code.

In this formulation, examples of non-messenger, functional components are those that preserve an accounting relationship for a non-messenger species of interest, such as in the aforementioned GPCR and PLCβ modules. Once the messenger species have been isolated in separate components, the rest of the module can be directly translated into functional CellML components. This is shown as the 'GPCR_Cycle' and 'PLC_Cycle' components in Fig. 2, which depicts the resultant partitioning of conceptual modules into CellML files and components following these principles.

The figure shows the contents of a main CellML file, which imports subsidiary files, each of which contain a model component and could have been developed independently by different researchers. The main file also contains the interface components, and the component that holds the parameters defining the cell's geometry. In Fig. 2, faded species represent those whose differential equation is contained by other components. Fluxes involving such species are exposed by their containing component, as they will be summed to a $J_{\text{source,species\_name}}$ variable in an interface component. There is no ligand nor PIP2 interface component as both species concentrations are fixed in those components. Two of the PLCβ forms, PLCβ-Ca^{2+}-G_α-GTP and PLCβ-Ca^{2+} are monitored by the PIP2 component (as $P\beta c$ and $P\beta cg$, respectively) which contains the reactions for their hydrolysis of PIP2.

If one allowed messenger molecules to also be the basis of an accounting relationship, one could make the claim that the G_α component could therefore contain all the forms of that species – which would include several receptor-bound and PLCβ-bound forms. Such a component would form a conservation cycle for G_α subunits as the fluxes for dG_d/dt, dR_g/dt, dR_{lg}/dt, dR_{lgp}/dt, dG_t/dt, dP_{cg}/dt and dP_g/dt (reactions R2–R13 inclusive) also sum up to zero. To do so, however, would force the removal of those forms from their previously assigned components, since although variable and parameter values can be shared between components, the differential equations which control the change of a variable over time can only belong to one. Therefore a choice must be made, in this case, whether to emphasise the accounting cycles of GPCR and PLCβ, or the accounting cycle of G_α. We assert that considering messenger molecules such as G_α as their own components rather than embodying them in an accounting cycle or chain: (1) resolves this conflict, (2) enables the translation of the GPCR functional module almost directly into a CellML component, keeping the semantic focus on the receptors as the basis of that functional unit and (3) allows easier connectivity with potential future components, as species likely to communicate between components (and therefore the nodes that communicate between the modules those components represent) are decoupled, requiring additions only to interface component code and not changes to existing functional or messenger components.

When formulating components, it may be realised that one conceptual module may become two or more components once the distinction between messenger and non-messenger molecules is made. For example, the 'IP3 Module' here becomes two components: (1) a 'PIP2' component centred on PIP2 and containing hydrolysis reactions forming the messenger IP3 (and Diacylglycerol, which is not shown for the purposes of this model) and (2) the IP3 component which is a messenger and contains its own degradation reaction. The 'IP3 Module' is, at the component level, a PIP2 hydrolysis component with an associated self-degrading messenger. The IP3 messenger molecule forms the communication carrier between the PIP2 component (and this model as a whole) and possible downstream components.

There is an important distinction that can be drawn between the IP3 degradation reaction (R16) and the G_α subunit self-hydrolysis reaction (R7). For the former, the reaction is known to be a lumped abstraction of the conversion of IP3 to either IP4 or IP2 [9]. This requires the interaction of other molecular species such as kinases and phosphatases, with their associated metabolic pathways. This is a clue that we have a potential functional component for this step. It is conceivable that a modeller wishing to use these components might wish to expand the IP3 degradation abstraction by including one or more of these pathways. To facilitate this, we advocate placing such abstractions that rely on other pathways in their own component (component 'IP3 degradation'), allowing the replacement of this reaction with a more detailed formulation without affecting the IP3 component itself or the rest of the model. By contrast, in the case of the G_α subunit self-hydrolysis, no other proteins or pathways are necessary for the reaction to occur, and hence this process

doi: 10.1049/iet-syb:20070020

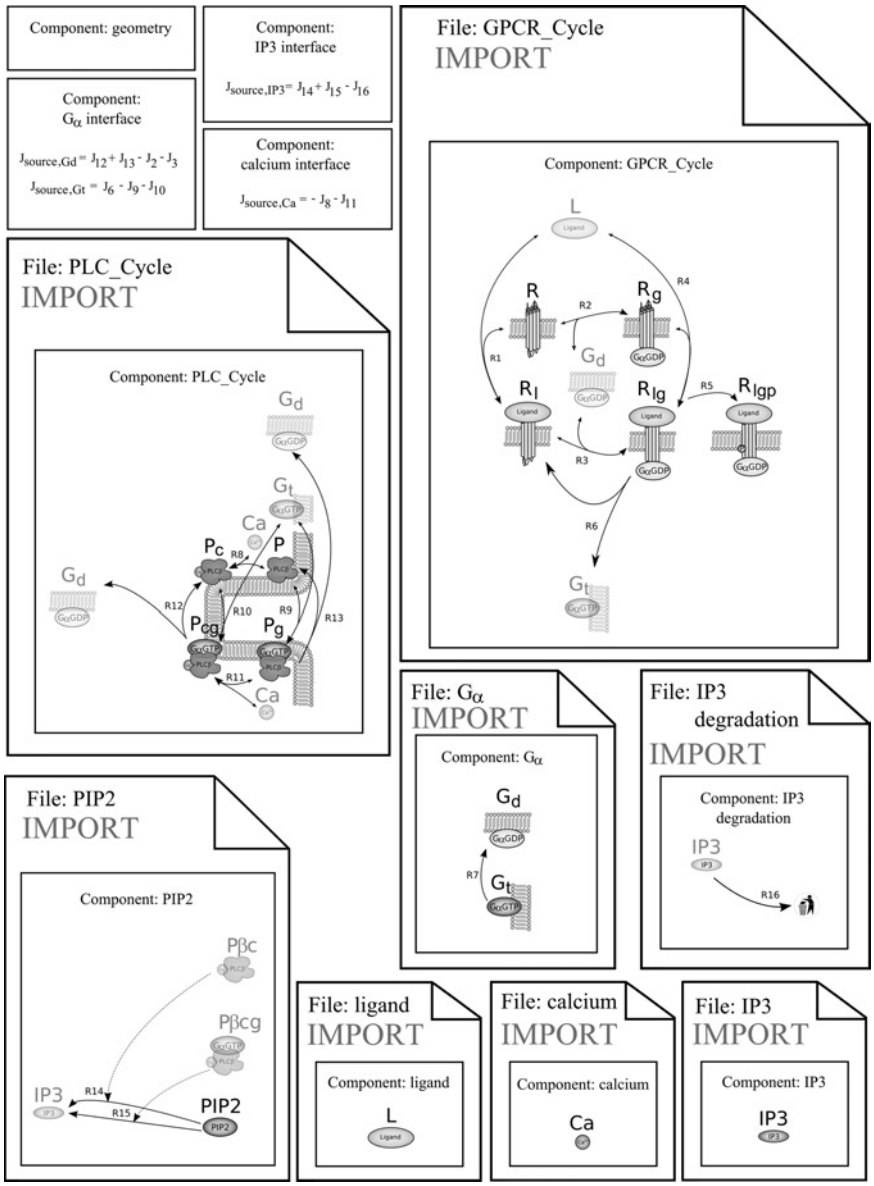

Figure 2 *Contents of the main CellML file for the model*

It contains the geometry component in order to define the cell, but imports signal transduction models from other files
Diffusive or mobile messengers are abstracted into their own components: 'ligand', 'calcium', 'G_α' and 'IP3'
Functional components 'GPCR_Cycle', 'PLC_Cycle', 'PIP2' and 'IP3 degradation' are also imported
Components which link via fluxes from other components are connected through interface components
Complete CellML code for the model can be found online (http://www.cellml.org/models)

is unlikely to be expanded upon further and can reside within the messenger component (component 'G_α').

4 Discussion

These components could be reused to construct models for other pathways. For example the GPCR component could be refitted and used for any instance of a GPCR which produces G_α-GTP – whether that product interacts with PLCβ or some other protein, in whatever pathway. Should for some reason a more detailed formulation for the GPCRs be needed for the IP3 pathway, the GPCR component can be replaced with one containing a more detailed formulation without affecting the other components. Additionally, by formulating the components where IP3 is in its own messenger component, we make it possible for components derived from additional functional modules to be added, read the IP3 signal and extend the pathway, without requiring code changes to the non-interface components in the existing model.

Aside from messenger against functional component distinctions as above, component boundaries may also be defined by physical or functional containment. For example, in a membrane which contains several ion channels, it seems logical that each ion channel should reside in its own

component, these components being added to larger models should those channels be required. This is the case for many existing electrophysiological models (see [12] for a range of electrophysiological examples coded in CellML).

In signal transduction, more complex relationships are possible including protein-to-protein interactions. For example, in this model PLCβ interacts with PIP2 to form IP3. This is implemented by the PIP2 component expecting PLCβ complexes and containing the kinetic rate constants for PIP2 hydrolysis by those enzymes. But the expectation of a specific complex from another component represents a tighter coupling than is ideal. In some cell types, PIP2 is hydrolysed by PLCγ complexes [13], which give different kinetic parameters for those reactions. The PIP2 component could be generalised (in keeping with its role as a PIP2 hydrolysis component) by the addition of reactions using those alternative isoforms, which when connected to the other components in the present model would have enzyme concentrations of zero and hence be inactive. These additional reactions have not been shown as they are not needed for this example.

Had the PIP2 and PLC_Cycle components been defined independently by different researchers, implementing only from their own modelling perspective, it is conceivable that both components could have been expected to control the differential equation for the PLCβ species. If both components were to be imported into the same model, the conflicts between two definitions of the same PLCβ species and possibly the contingent reactions would have to be resolved. The current CellML specification (version 1.1) does not provide a standardised way of handling this potential conflict, but work is underway to address these concerns. Future plans for CellML include the binding of biologically-relevant identifiers to CellML elements via metadata tags that use the biological pathways exchange language BioPAX (Biological Pathways Exchange - http://www.biopax.org). Two CellML variables or reactions with the same identifier will be semantically equivalent regardless of their CellML representation, and it is envisaged that duplicates could be automatically determined during model construction. Once this functionality is available, it is likely that the 'leading practice' for CellML model design will include defining each species and reaction in its own component, which is then encapsulated into higher level components to represent the physical or functional modules as defined above. This design enables the reformulation of higher level components by the model builder at the time of model aggregation. Duplicates once detected could have the resultant conflict resolved by manually choosing an alternative whose components are then linked (at the CellML code level) automatically. Thus for signal transduction, modules may still be combined without having to edit the code for the components that implement them, but we envisage that a strictly black-box approach may not always be possible, as component contents would have to be understood well enough for the modeller to decide how such conflicts should be resolved. This technology would support rather than replace human intelligence in the model building process.

5 Glossary

Accounting chain/cycle: A set of reactions, which accounts for the mass of a particular molecular species. It may include conservation of mass but may not (for example, where species are removed from the model). Depending on the reactions, possible topologies of the graph of this reaction set include cycles and chains.

Component: In general terms; a distinct unit of a model implemented in some computer readable format. When used specifically in the context of the IP3 model in this work, components are CellML 'components' as defined in the CellML specification (available at http://www.cellml.org). We distinguish between functional and messenger components.

Functional component: A component that implements a functional module once messenger species are separated out into their own messenger components.

Messenger component: A component that encapsulates the amount of a molecular species that is likely to be used as a messenger species between functional components.

Module: A conceptual entity encompassing a biological function. In our work, a module forms an accounting chain or cycle for the molecular species deemed key to that biological function. Modules are implemented in computer readable form by functional and messenger components.

6 Acknowledgments

The authors wish to thank David P. Nickerson, Matthew Halstead and Poul Nielsen for useful discussions. MTC was supported by a Bright Futures Doctoral Scholarship from the Foundation for Research, Science and Technology of New Zealand. EJC and PH are grateful for support from the Wellcome Trust and a Maurice Wilkins Center grant from the New Zealand Tertiary Education Committee.

7 References

[1] KITANO H.: 'Systems biology: a brief overview', *Science*, 2002, **295**, pp. 1662–1664

[2] ALON U.: 'Biological networks: the tinkerer as an engineer', *Science*, 2003, **301**, pp. 1866–1867

[3] HUNTER P.J., BORG T.K.: 'Integration from proteins to organs: the Physiome Project', *Nat. Rev. Mol. Cell Biol.*, 2003, **4**, pp. 237–243

doi: 10.1049/iet-syb:20070020

[4] LE NOVÈRE N., FINNEY A., HUCKA M., *ET AL.*: 'Minimum information requested in the annotation of biochemical models (MIRIAM)', *Nat. Biotechnol.*, 2005, **23**, pp. 1509–1515

[5] BHALLA U.S., IYENGAR R.: 'Emergent properties of networks of biological signaling pathways', *Science*, 1999, **283**, pp. 381–387

[6] LUKAS T.J.: 'A signal transduction pathway model prototype i: from agonist to cellular endpoint', *Biophys. J.*, 2004, **87**, pp. 1406–1416

[7] HARTWELL L.H., HOPFIELD J.J., LEIBLER S., MURRAY A.W.: 'From molecular to modular cell biology', *Nature*, 1999, **402**, pp. C47–C52

[8] LLOYD C.M., HALSTEAD M.D.B., NIELSEN P.F.: 'CellML: its future, present and past', *Progr. Biophys. Mol. Biol.*, 2004, **85**, pp. 433–450

[9] COOLING M., HUNTER P., CRAMPIN E.J.: 'Modelling hypertrophic IP3 transients in the cardiac myocyte', *Biophys. J.*, 2007, **93**, (10), pp. 3421–3433

[10] ALON U.: 'An Introduction to systems biology: design principles of biological circuits' (Chapman and Hall/CRC, Boca Raton, 2007)

[11] NICKERSON D.P., NASH M.P., NIELSEN P., SMITH N., HUNTER P.: 'Computational multiscale modeling in the IUPS physiome project: Modeling cardiac electromechanics', *IBM J. Res. Dev.*, 2006, **50**, pp. 617–630

[12] NICKERSON D.P., HUNTER P.J.: 'The Noble cardiac ventricular electrophysiology models in CellML', *Prog. Biophys. Mol. Biol.*, 2006, **90**, pp. 346–359

[13] VAN LEEUWEN J.E.M., SAMELSON L.E.: 'T cell antigen-receptor signal transduction', *Curr. Opin. Immunol.*, 1999, **11**, pp. 242–248