Cytoview: Development of a cell modelling framework

Prashant Khodade[1], Samta Malhotra[2], Nirmal Kumar[2], M Sriram Iyengar[3], N Balakrishnan[1] and Nagasuma Chandra[1,2,*]

[1]*Supercomputer Education and Research Centre and* [2]*Bioinformatics Centre, Indian Institute of Science, Bangalore 560012, India*

[3]*University of Texas Health Science Center, Houston, Texas 77030, USA*

**Corresponding author (Fax, 91-80-23600551; Email, nchandra@serc.iisc.ernet.in)*

The biological cell, a natural self-contained unit of prime biological importance, is an enormously complex machine that can be understood at many levels. A higher-level perspective of the entire cell requires integration of various features into coherent, biologically meaningful descriptions. There are some efforts to model cells based on their genome, proteome or metabolome descriptions. However, there are no established methods as yet to describe cell morphologies, capture similarities and differences between different cells or between healthy and disease states. Here we report a framework to model various aspects of a cell and integrate knowledge encoded at different levels of abstraction, with cell morphologies at one end to atomic structures at the other. The different issues that have been addressed are ontologies, feature description and model building. The framework describes dotted representations and tree data structures to integrate diverse pieces of data and parametric models enabling size, shape and location descriptions. The framework serves as a first step in integrating different levels of data available for a biological cell and has the potential to lead to development of computational models in our pursuit to model cell structure and function, from which several applications can flow out.

[Khodade P, Malhotra S, Kumar N, Iyengar M S, Balakrishnan N and Chandra N 2007 Cytoview: Development of a cell modelling framework; *J. Biosci.* **32** 965–977]

1. Introduction

It is becoming increasingly clear that *in silico* modelling of biological systems is a far more complex endeavor than previously imagined (Tomita *et al* 1999; Ideker *et al* 2001; Loew 2002; Hunter and Borg 2003; Thomaseth 2003; Kiehl *et al* 2004) mainly because the complexity of biological systems is not amenable to easy, simplistic solutions. In this respect, the biological cell is a natural self-contained unit, of prime importance. The fundamental unit of living tissue, in fact of life itself, is the biological cell. Currently there is enormous interest in *in silico* modelling of the cell in its many aspects. The cell is, of course, an enormously complex machine which can be understood at many levels, functional, signaling, metabolic, and regulatory and so on. However, there is a growing recognition that understanding its structure and the physical nature of intracellular objects, as well as their three dimensional spatial relationships, can yield significant insights into physiology and functionality (Thomaseth 2003, Hunter 2004). Here we report a framework to represent the various aspects of a cell, which is an important aspect of cell modelling.

Although cell modelling in its various aspects is a subject of intense study currently across the globe (Loew 2002, *http://www.nrcam.uchc.edu*, Tomita *et al* 1999, Hunter & Borg 2003, Hunter 2004, Rosse & Mejino Jr, 2003), several questions remain open, warranting further work in this area. One main lacuna is the lack of integrated models that span across cell morphologies to organelle structure, function and dynamics relating ultimately to gene or protein level

Keywords. Cell modelling; cell morphology; cell ontology; computational models; dotted representation

Abbreviations used: DAG, Directed acyclic graph; RER, rough endoplasmic reticulum; VRML, virtual reality modelling language

knowledge. Here we seek to address this issue and have worked towards a framework for such integration, with an emphasis on the cell morphological structures to start with.

2. Results and discussion

2.1 *Overview of the framework*

The main objectives of this framework termed Cytoview have been to develop methods to enable development of computational models to describe biological cells, incorporating information at different levels of hierarchy. Different aspects required to be addressed in the framework are (i) a systematic vocabulary to describe individual sub-cellular structures as well as their inter-relationships, (ii) precise parameters to define and describe the individual substructures and features, (iii) a model to integrate the individual sub-structures using their parameters into a whole cell and (iv) ability to apply these models in seeking answers for specific biological questions. These issues have been dealt with in the framework through the use of (i) ontologies, (ii) feature extraction and (iii) model building. An example application is also described to illustrate the usefulness of the framework. A schematic diagram that has been used in creating the framework is shown in figure 1.

2.2 *Cell ontology*

One of the important challenges in integrated biology is the formal description of phenotypic data and their correlation to the relevant genotypic data. While this type of correlation is well understood and highly developed in

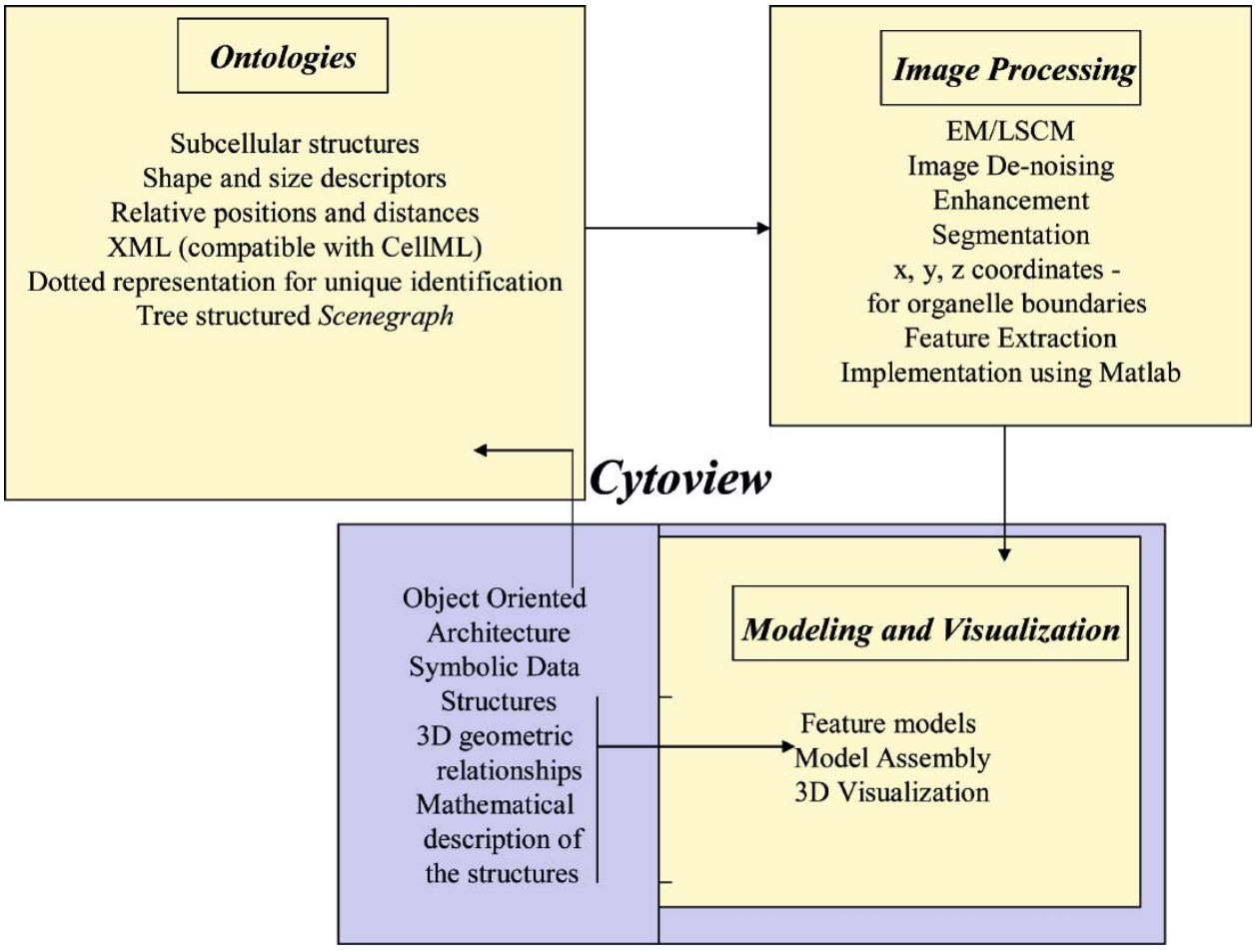

Figure 1. An overview of the framework showing the use of ontologies, parametric models and the modelling and visualization of whole cells. Important aspects in each panel are listed. Image processing panel (discussed in detail elsewhere) can provide precise definitions of various parameters from real biological images.

some cases (e.g. the cause of sickling of red blood cells), there is hardly any information available in a majority of situations. It is important to bring the available data at any level of hierarchy (e.g. electron microscopic images of cells, metabolic capabilities of these cells, disease associations or even specific gene or protein level disorders, which can be studied at the level of their three-dimensional structures) into well-structured computationally tractable representations. A formal scheme to structure the diverse data using standard vocabulary is therefore very important. In order to obtain the various parameters required for modelling the cell, structured information of each of the cell types as well as their individual components is required.

An ideal ontology should enable the mapping of data at various levels of hierarchy. Although detailed information at every level is not as yet easily available for a given biological system, the schema should provide for easy integration of data as and when it becomes available. Taking the cell as the layer of focus, we can propagate both deeper and also higher-up in the organizational hierarchy. Figure 2 illustrates the different levels of hierarchy considered here. Further, enormous amount of knowledge has been gathered in the literature for individual systems or processes at different levels, representing data about a certain process in different levels of abstraction. Depending upon the question being addressed, it may be sufficient to analyse data at one or two levels of abstraction only. However, to get a broader perspective of the cell and how seemingly disparate processes influence other processes in the cell, an integration of the whole and cross mapping between different levels is important. Figure 2 also lists the models available at different levels and what each of them encodes as well as the methods to study such models. Given that biological systems and processes are understood at many different levels and in many different aspects, it comes as no surprise that many different kinds of models should exist in practice. It is important to understand the abstraction levels of the models, so that conclusions are drawn at appropriate levels from the analyses.

While there is enormous diversity in cell types, structures and functions, they are all based on fundamentally common themes. An attempt has been made to describe various cell types into structured ontologies by Ashburner and coworkers (Bard *et al* 2005), which is used as a starting point here. It includes about 680 different types of cells with all components present inside them. The ontology

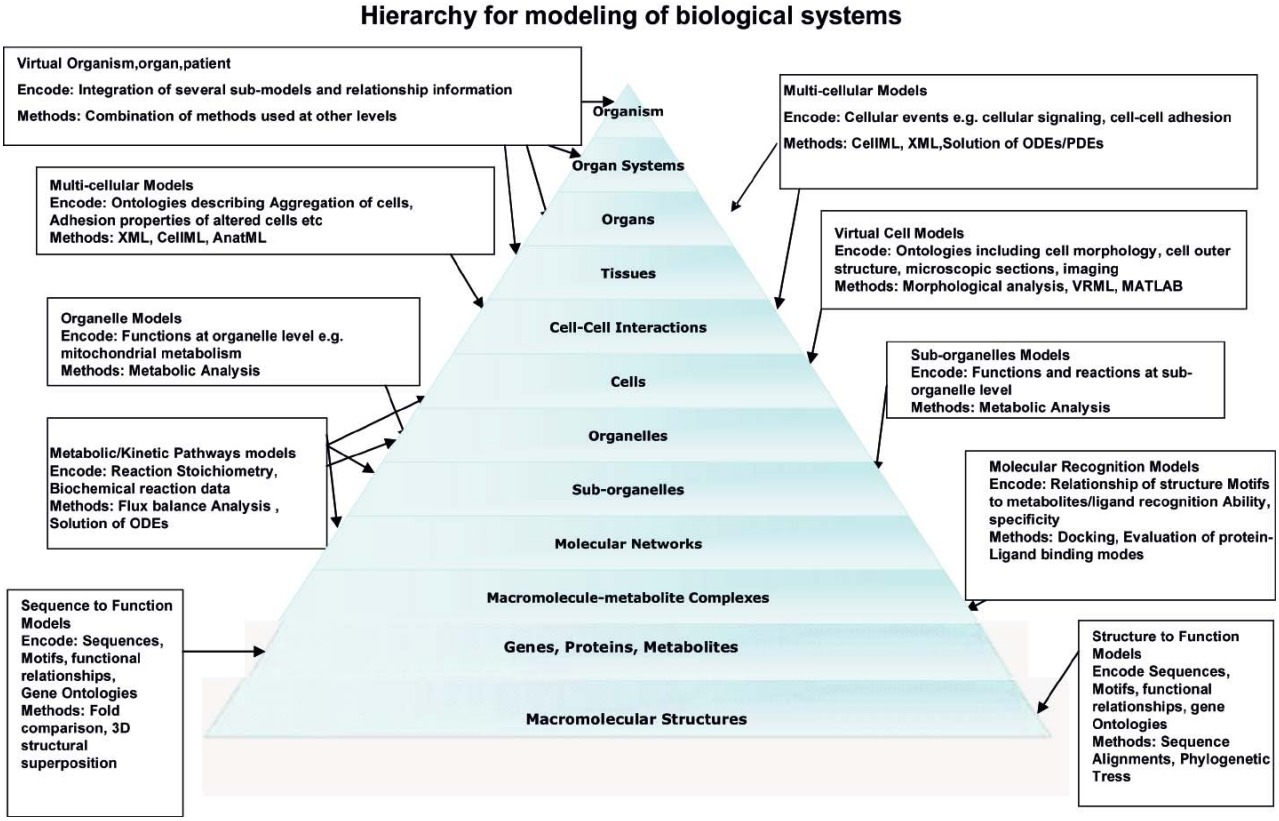

Figure 2. Hierarchy in modelling biological systems. Different models available at various levels are displayed as different levels in prism. We divided the hierarchical structure into 2 levels; cell being the layer of intersection between the two.

developed previously (Bard *et al* 2005) consists of concepts or terms (nodes) that are linked by two types of relationships (edges), which means that the ontology appears as a complex hierarchy (which essentially is_a directed acyclic graph, or DAG) where a given term or concept may not only have several children, but also several parents. The parent and child terms are connected to each other by is_a and develops_from relationships. The former is a subsumption relationship, in which the child term is a more restrictive concept than its parent (thus chondrocyte is_a mesenchymal cell). The latter is used to code developmental lineage relationships between concepts, for example that a hepatocyte develops_from a mesenchymal cell. The is_a relationship implies inheritance, so that any its children inherit properties of the parent concept; the develops_from concept carries no inheritance implications.

While this provides an excellent handle to study cell classification and cell type relationships, newer vocabulary is required to describe a single cell itself with all its sub-cellular structures. Further, this vocabulary should pave way for integrating cell morphologies with functional information of the cell as a whole or a set of molecules within the cellular or molecular level detail of the individual proteins, genes, and their regulation. To cater for these requirements, here we propose the use of a tree based approach rather than DAG based approach. Also we incorporate the function information in each node of our cell tree so that we get not only structural but also functional information. Our representation of cell ontology therefore leads a further step ahead, representing the cell in much more detail.

2.2.1 *The tree structure:* In computer science, a tree, which is also a special case of a graph, is a widely used data structure that emulates a tree structure with a set of linked nodes (Cormen *et al* 2001). Each node has zero or more child nodes, which are below it in the tree. A node that has a child is called the child's parent or ancestor node. A node has at most one parent. This is the most significant difference between DAG approach and our tree-based approach. The topmost node in a tree is called the root node. It is the node at which all operations on the tree begin. All other nodes can be reached from it by following edges or links.

The cell can be considered as a system having components and subcomponents from subcellular organelles to proteins (figure 3). Here we arrange these components at various levels of hierarchy. The root of the tree begins at the cell. At the next level we have types of cells such as nervous, muscular and immune etc, and at subsequent level we have organelles. The concept of layered representation is similar to *scenegraph* concept in computer graphics. Scenegraphs are a collection of nodes in a graph or tree structure. This means that a node may have many children but often only a single parent, the effect of a parent is apparent to all its child nodes. An operation applied to a group automatically propagates its effect to all its members. A geometrical transformation matrix is associated at each group level.

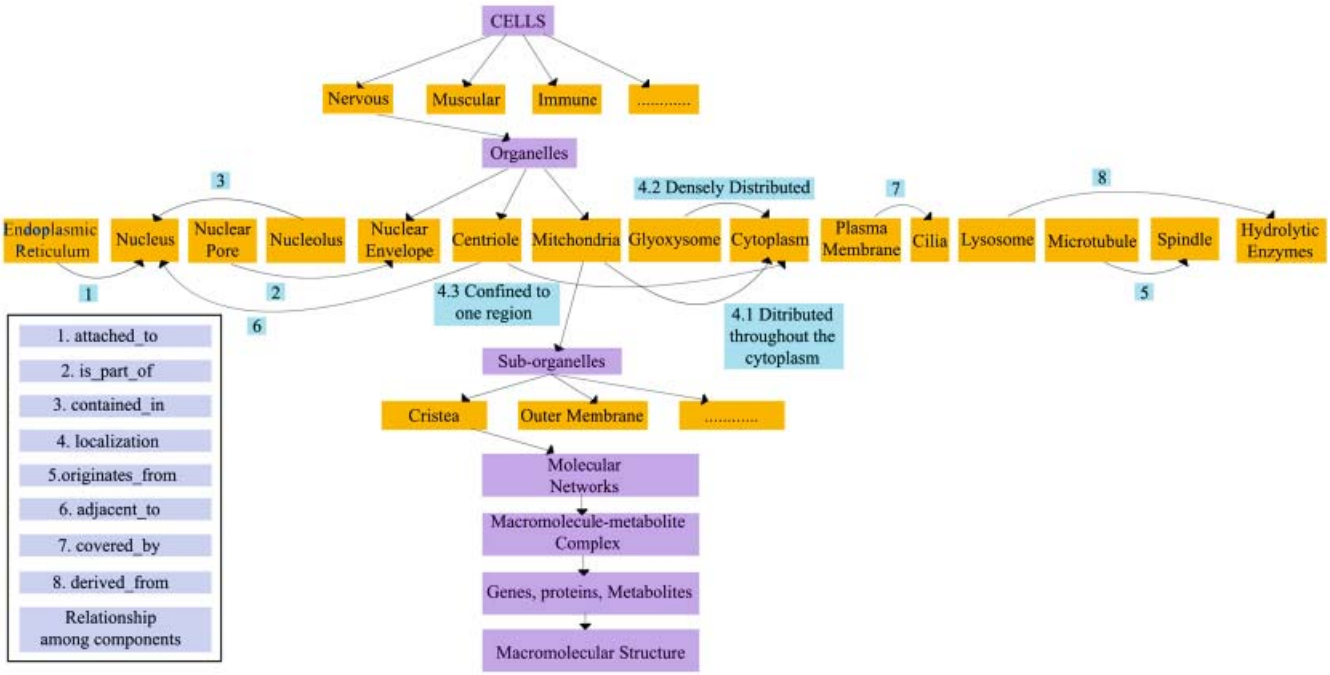

Figure 3. Improvised cell tree data-structure to store information about the cell. Application of scenegraph for cell modelling; the concepts of scenegraph and topological information lead us to extracting the substructure location information from cell images.

A common feature is the ability to group related shapes/objects into a compound object which can then be moved, transformed, selected etc. as easily as a single object.

The concept is useful not only to model the cellular structure, but it also allows manipulation of the cell. If we manipulate a given node its environment also needs to get manipulated. Thus if we mutate a protein, the semantics used for modelling the cell should cater to reflecting the change in some parameter of the subcellular structure, where such mapping is possible. The individual proteins/genes in the cell are represented as the leaf components of the tree. One can imagine the enormity of the tree. Although there are several cell types and several types of subcellular components as well as high diversity in the proteome component, using the tree structure for representation does not make the navigation or retrieval slow. The tree is constructed such that its height is not more than a certain number of layers. When number of instances in each layer is increased, the tree merely grows in its width but not in its height.

2.2.2 *Dotted representation:* In the dotted representation we use a concept from computer networking, where the numbers separated by dots represent Internet protocol (IP) addresses. This simple scheme can generate unique IDs necessary for the growing number of computers in the world. A similar scheme applied to cells can be illustrated as: when an example of a protein present in more than one type of cell is considered, the need to differentiate the same protein between the first and the second cell arises. We use a dotted representation to make them unique. In this representation the numbers are separated by dots. The last number reflects the most detailed node.

In the dotted representation we organize the information hierarchically, each hierarchy being separated by a dot in the notation. We consider here an example of dotted representation. Consider a protein present in the cytoplasm with primary accession number P19367 and a PDB: 1HKB (Aleshin *et al* 1998). We describe this protein as 10.7.0.8.3.3.2, as illustrated in figure 4. In situations, where information on the cell state is available, the dotted representation can be prefixed with a letter or a number indicating one of the given states of the cell. In the above representation, the number 10 represents the type of cell, where as number 7 describes the cellular organelle, which in this case is the cytoplasm. The number 0 tells that no sub-component is present in corresponding hierarchy. The number 8 gives the pathway information describing that the protein is involved in glycolytic pathway. The number 3 tells that the protein forms the complex with Glu-6-phosphate. The number 3 represents the protein with Uniprot id P19367. Thus we can get the information about the protein. Further if we want to have more fine grained information about the three dimensional structure we traverse to next level of tree, in this particular case this has been denoted by number 2, which stands for secondary structure. We arrange this cell hierarchy in the form of a tree as shown in figure 4. The root node of the tree is the cell. Further branches specify types upto the protein structural level. Supplementary table shows the organelles in a typical cell, which is encoded in a structured way.

Handling complexities and special cases: We have various complexities to take into c\onsideration when we consider tree data structure for the cell. Some components may not be present in some cells leading to null fields in the representation, e.g. in the case of mature red blood cells (erythrocytes) in humans. These are the simplest and most important cells and to accommodate them in the representation, we should have unique representation and yet an efficient implementation. Two ways of handling this issue are to either keep some subcomponents as null or to move some components up in the hierarchy. The solution we follow is to move some subcomponents up in the hierarchy and keep count of number of levels, the node is occupying. This information is kept in the higher nodes. Also a rule file is created in each case for collapsing and expanding the tree as needed. Thus we have for red blood cells count related to levels it is occupying. When we require information related to haemoglobin but structural information is not required, we start at the root node, cell, down the hierarchy we have next, cell type, which is red blood cell (let us say, its representation is 7) and further down haemoglobin also represented by 7. The haemoglobin is represented with 7.0.0.15.5.7.0. We start at the root node and come to red blood cell, which has id 7. Now we stay at the red blood cell node until non-zero number comes in the dotted representation. This is shown by an arrow, which comes to itself as we travel. When non-zero number 15 comes, we travel down which brings us to oxygen transport node. Further down we travel to node 5, which is a complex of haemoglobin with oxygen, and then to haemoglobin represented by id 7. As we do not need more refined information about structure we stay at haemoglobin node, which has the required information. This gives us an efficient implementation while making the representation uniform. Users will not have to be aware of the number of levels present in a particular cell.

2.2.3 *Relationships among cellular components:* The cell ontologies have defined relationships of one cell with respect to the other, where applicable. Besides the evolutionary relationships, the OBO cell ontology (Bard *et al* 2005) also links different cells through the keywords in their annotations referring to their location by tissue or organ or by function.

In addition to these, we require definitions of sub-cellular structures within a given cell, in order to model its morphological and spatial relationships. Here we propose the use of four types of relationships, as detailed in table 1.

Figure 4. Representation of hexokinase present in the cytoplasm of neuroendocrine cells. In special cases such as when a null field is encountered in any layer, the pointer will stay at same place till it finds next non-zero number.

While the major relationships are listed here, it is possible to define more such relationships, which can be added easily, as they become available. The spatial relationships specify the location of each of the sub-cellular structure within the cell. For n defined components in the cell, a $n \times n$ matrix is prepared as shown in Table 2. The size relationships can be defined either qualitatively or quantitatively, if precise parameters are available. An example is, the size of the lysosome (200 nm) is similar to that of the centriole (200 nm) in the same cell. Functional and chemical relationships can also be defined when explicit data become available either for components within the cell or for defining inter-cellular interactions. Some examples are the gradient of a given molecule within the cell or interaction of a protein from cell A with another of cell B (e.g. during phagocytosis)

In several cases such as those from cellular imaging, we have information about the relative spatial location of the sub-structures in the cell. In such situations, it is also possible to define distance relationships among different sub-structures, as shown in table 3. The centroid of each sub-structure is computed and distances between each pair of them are calculated, which provides an idea about the relative position of each of the sub-structures. These distances in turn can be used as restraints during whole cell assembly and further in simulations based on whole cell models.

Table 1. List of relationships between cellular components. Each of them is given an index key, which will be used in the relationship matrices

Type of relationships		Description/ examples
Spatial		
1.	attached_to	Endoplasmic reticulum is attached to the nucleus.
2.	is_part of	Nuclear pore is part of the nuclear envelope.
3.	contained_in	Nucleolus is contained in the nucleus.
4.	Localization	(4.1) distributed throughout, as in mitochondria distributed throughout cytoplasm, (4.2) local regions with high concentration, example: glyoxysomes, (4.3) confined to one or two regions only, example: Centriole is confined to one region in cytoplasm.
5.	originates_from	When a component gets significantly modified into another distinct component, the latter is said to originate from the former. eg., Spindles originate from microtubules.
6.	adjacent_to	Centrioles are adjacent to the nucleus.
7.	covered_by	Plasma membrane is covered by cilia.
8.	derived_from	When a smaller component is not structurally a distinct part of another larger component, but can be separated from it (with or without minor modifications), it can be said to be derived from the larger component, e.g. many hydrolytic enzymes are derived from lysosomes.
Size		
9.	Quantitative	When a component is defined with a specific size: eg., typical size (diameter) of a lysosome is in the range of 200-500 nm
10.	Qualitative	When the sizes of two components are described in relative terms; eg., nucleolus is always smaller than that of the nucleus
Functional classes		
11.	Metabolic	Functional role of the mitochondria
12.	Transport	Pore size of the cell membrane correlated with the size of the molecule it has to transport.
13.	Support	Functional role of the microfilament.
14.	Reproduction	Centriole replication and division of centrosome formation.
15.	Storage	Storage of nutrients and waste products in vacuoles in plant cells.
Chemical		
16.	Gene expression	
17.	Metabolite recognition	
18.	Enzymatic activity, etc.	

2.3 *XML representation*

To make use of the representation and data structure suggested above there is need to address storage issues as well. The information stored should be understandable to human and it should be interpretable by the machine. The above dotted representation can directly map to XML (*http://www.w3.org/XML/,*) syntax. The example for Hexokinase when stored in XML, which is also compatible with CellML, is shown in figure 5.

This representation is compatible with current CellML (Lloyd *et al* 2004), which makes use of FieldML (*http://www.cmiss.org/openCMISS/wiki/FieldMLConcepts*), MathML (*http://www.w3.org/Math/*) and AnatML (*http://www.physiome.org.nz/anatml/pages/index.html*) formats. We can have in single XML file data about cell representation giving information regarding pathways, kinetic parameters and geometries. Currently we have small dataset of the information about different cell types, their geometries, their components and sub-components. There is need to do large scale web-mining for the information about cell types and their components, subcomponents and proteins they contain. Such a database can make use of the above methods. All this information can be linked to gene ontologies (Ashburner *et al* 2000). For the success of systems biology endeavors, such a database will be useful and success of such a database depends on the methodologies suggested in this framework.

Table 2. Relationship matrix illustrating the spatial relations of each of the components in a cell with respect to all the others

Name of Organelle	Centriole	Chloroplast	Endoplasmic reticulum	Golgi apparatus	Lysosomes	Mitochondria	Ribosomes	Vacuoles	Nucleus	Cilium	Glyoxysome	Hydrogenosome	Nucleolus	Peroxisome	Cell membrane	Flagella	Cytoplasm
Mitochondria	0	0	0	0	0	0	3	0	6	0	0	0	0	0	0	0	4.1
Chloroplast	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	4.2
Centriole	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	4.3
Cilium	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0
Golgi Apparatus	0	0	0	0	0	0	0	5	6	0	0	0	0	0	0	0	4.2
Cell Membrane	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	7	0
Vacuoles	0	0	0	5	5	0	0	0	0	0	0	0	0	0	0	0	4.1
Nucleolus	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0
Ribosome	0	3	7	0	0	3	0	0	0	0	0	0	0	0	0	0	4.1
Peroxisome	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	4.1
Nucleus	6	0	2	6	0	0	6	0	0	0	0	0	3	0	0	0	4.3
Lysosome	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	4.1
Glyoxysome	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4.2
Hydrogenosome	0	0	0	0	0	8	3	0	0	0	0	0	0	0	0	0	4.1
Flagella	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0
Endoplasmic Reticulum	0	0	0	0	0	0	7	0	1	0	0	0	0	0	0	0	4.2

Zero refers to a null relationship. Other numbers are index keys defining the type of relationship while the last column indicates the location of components in cytoplasm. (4.1) Distributed throughout cytoplasm, (4.2) regions with high density and (4.3) confined to one or two particular region(s) within cytoplasm.

Table 3. An example distance matrix indicating distances (nm) between centroids of different sub-structures in a typical animal cell

	Mitochondria	Nucleus	Nucleolus	Lysosome	Vacuole	Entire cell	Golgi complex
Mitochondria	0	129.92	111.14	56.31	63.72	134.96	126.14
Nucleus	129.92	0	23.5134	137.92	153.19	52.32	80.41
Nucleolus	111.14	23.51	0	126.95	130.06	41.47	88.68
Lysosome	56.31	137.92	126.95	0	119.93	162.00	99.18
Vacuole	63.72	153.19	130.06	119.93	0	134.70	177.22
Entire cell	134.96	52.32	41.47	162.01	134.70	0	129.30
Golgi complex	126.14	80.41	88.68	99.18	177.22	129.30	0

2.4 *Parametric models and Feature extraction*

Much of the information on cellular structures at various levels of detail that we have today has been obtained from different types of cellular imaging techniques. The most prominent of these techniques are electron microscopy and its variants as well as fluorescence microscopy. Converting the qualitative data into quantitative type is required to model the cell. Image processing and computer vision techniques help to convert the qualitative information in cell images into quantitative information using the features extracted and use them in the design of mathematical models. Following image pre-processing to remove noise and identify objects, image enhancement and morphological operations carried

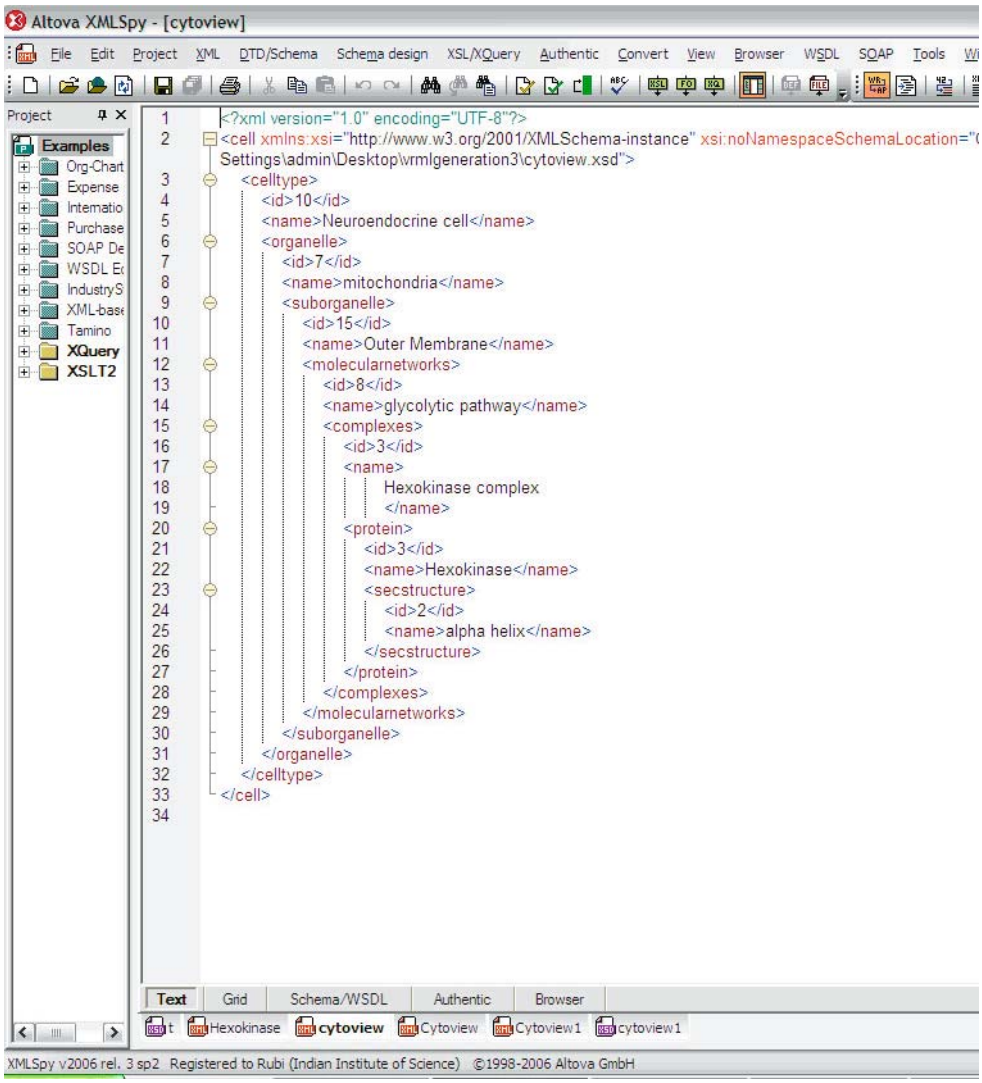

Figure 5. Screenshot of the XML representation of hexokinase using XMLSpy (*http://www.altova.com*)

out to further reduce noise (algorithmic details of this will be published elsewhere, as they are outside the scope of this article) and segmentations yield outer boundaries of the cell as well as those of the sub-cellular organelles. Several features such as area, eccentricity, major axis, minor axis, orientation and Euler number were extracted for each of the images which help in describing sub-cellular organelles and in obtaining morphological models of cells.

2.5 *Algorithms for representation and assembly of cellular and intracellular structures into coherent assemblies and their visualization*

As a result of processing of the images we get cell substructure boundaries. These boundaries are used to reconstruct the substructure. Currently, this has been applied to 2D electron microscopic images. With the application of cryo-electron tomography for cells this methodology can be applied to substructure reconstruction for 3D images as well. From the centroids of the cell and its substructures extracted as a feature, spatial distances were calculated between cell organelles and assembled into a whole cell model, with application of appropriate transformations. For example some parts of rough endoplasmic reticulum (RER) image also provide some information about the lysosome. Similarly, image containing the lysosome may provide some clues about the RER. These clues when combined, provide spatial information of the two components in the form of geometric constraints, and thus used in modelling their assembly. The problem of obtaining an optimal placement of the organelles such that distance constraints are satisfied,

are amenable to well-established optimization protocols, which can be integrated easily. A sample toolkit developed on based of this concept is available at *http://nscdb1.bic.physics.iisc.ernet.in/cgi-bin/cytoview/input.cgi*.

The next task in the cell modelling process is to visualize the structured information thus generated. Earlier work related to model visualization has been described in (Kremer *et al* 1996). For visualization of the data generated after image processing we used virtual reality modelling language (VRML). Rendering and interactive visualization provided by VRML is compatible with CellML. VRML has been used not only to enable 3D visualization of cells, but also to represent the information with minimum amount of data still representing it to the maximum extent, so that the information extracted, becomes accessible to the community over the Internet. VRML is thus the choice of modelling language and also makes visualization easier through the browser with an appropriate plugin. Results obtained for cell organelles and their composition into a complete cell models is shown in figure 6.

Due to lack of easy availability of 3D images, we used 2D images and generated an approximate 3D model by generating surface of revolution. While generating models from 3D images would be ideal for capturing the morphology accurately, approximate models such as these where 2D data is extended to 3D with the help of indirectly derived depth parameters serve to demonstrate the capability of the framework in generating and manipulating cellular models. The extrusion method from the VRML modelling is used for generating an approximate surface of revolution. Setting the spine along y-axis and making the scale vary generates the surface of revolution. The 3D object generated this way can be seen in the browser having VRML plug-in. Thus given an image of the cell or a collection of images of subcellular organelles, 3D models of the cell organelle were generated, which were then viewed and manipulated using VRML. As a result we get an interactive 3D whole cell model, which is an assembly of individual cell organelles. In VRML terms, figure 6 illustrates a scenegraph of the cell.

The model has been composed with Internet scene assembler from Parallel Graphics (*http://www.parallelgraphics.com/*). The model has low storage requirement as we have currently only cross-sectional information about the cell and organelles and the 3D information comes implicitly in the form of extrusion. The model requires currently runtime load due to the fact that the required coordinates are calculated at runtime. With the availability of the image slices for the 3D model, we can generate a more realistic model. In such cases, we will have image slices and hence corresponding cross-sections for each object contained, from which we can generate 3D models. Appropriate encoding of texture of the cell surface can further refine these models.

2.6 *Application scenarios and future prospects*

Computational modelling of biological systems can achieve integration along several dimensions. Structural integration across physical scales of biological organization ranges from genes to cell, tissue, organ, and whole organism where as functional integration, involves comprehension of data related to gene expression, protein synthesis, signal transduction, metabolism, ionic fluxes, cell motility and many such functions (Ideker *et al* 2001). The challenges

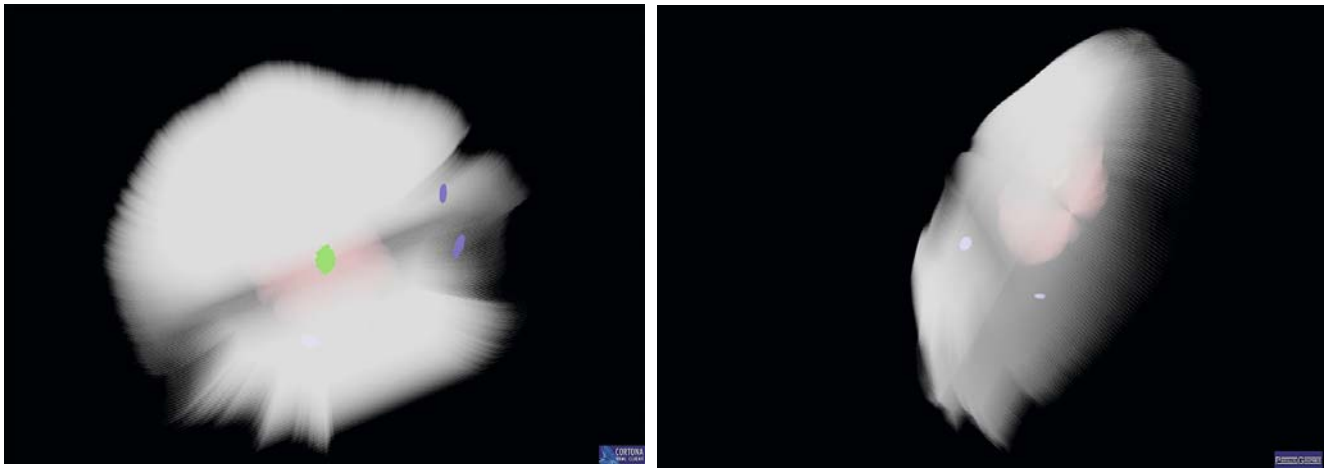

Figure 6. VRML-output shows the snapshots of a 3D model from a cross section of the entire cell and subcellular organelles. The left figure shows an open part of entire cell with organelles embedded in it. The right figure shows the same cell in an orthogonal view. The subcellular organelles – mitochondria, nucleus and nucleolus can also be seen.

of structurally integrated and functionally integrated computational modelling tend to be different. Functionally integrated biological modelling is a central goal of what is now being called systems biology (Ideker *et al* 2001). However structural integration is more complex than the functional integration. There exists relation between them also, which can be taken as horizontal integration between structure and function. The object oriented implementation of tree structure and creation of comprehensive cell database with web-mining techniques is one way of achieving this integration.

Just as software systems like AutoCAD represent the physical structure of engineered objects such as automobiles and aircraft, a method that will enable the knowledge representation of the morphology of biological cells will be useful. Although enormous quantities of data are available in the literature, they are not readily usable by the researchers in the broader domain, because each piece of data has been generated with a specific but narrow focus, as is often the case with experimental biology, and do not automatically render itself to be integrated with wider perspectives and therefore do not always benefit from parallel advances in other areas. Further, the data have been obtained by disparate researchers and methods in vastly different contexts and are not also in computer-accessible formats. There is currently no integrated, codified representation. Rather than creating such a representation on paper, modern techniques of computer science have emerged which will enable this knowledge to be integrated and made available in an interactive form enabling myriad applications, including visualization and a framework for development of advanced software, to model complex aspects of cell functioning and cell-cell interaction.

Some of the obvious benefits of the Cytoview that can be envisaged are: (i) as a framework and tool for precisely representing cell types, sub-types and their components, (ii) as an useful tool for classifying and clustering various cell types, (iii) as a starting point for studying cells from systems biology perspectives both scaling-up to larger tissues and organs etc. as well as scaling down to individual molecules level should be possible, and (iv) as a basis for studying diseases and the progression of disease in complex cell assemblies, where the physical structures have been altered due to disease states, where-in they can form the basis for diagnosis too. As an example of representing different cell types and further using them for classification, granulocytes such as eosinophils, basophils, neutrophils can be distinguished from agranulocytes such as lymphocytes, monocytes or erythrocytes by the presence of granules inside the cell, as can be readily deduced from the various descriptors (described in the accompanying article) used in modelling these cells. Further as an example of studying complex cell assemblies, models of eosinophils interacting with mast cells both in healthy states as well as in disease states such as chronic allergies, can be generated and compared with each other. Finally, the example below describes how this framework can serve as a starting point for integrating knowledge at different levels of hierarchy. A protein ACTHr, a G-protein coupled receptor, also known as MC2R in the annotated databases (NCBI *http://www.ncbi.nlm.nih.gov*; ExPASy Proteomics Server, *http://www.expasy.org*; GeneCard *http://www.genecards.org*), the expression of whose gene is upregulated by its own ligand, ACTH, via a cAMP-dependent pathway (Mesiano *et al* 1996), binds to the hormone ACTH and triggers a signal transduction pathway to result in stimulation of steroidogenesis (Mountjoy *et al* 1992), apart from activating the adenylate cyclase pathway and consecutive activation of protein kinase A (Bourdeau and Stratakis 2002). This protein is mainly expressed in the adrenal cortex (Beuschlein *et al* 2001), but has also been identified in the human skin (Solminski *et al* 2004) and has been linked to diseases such as adenoma (Zwermann *et al* 2004), ovarian steroid cell tumor (Lin *et al* 2000), Cushing syndrome, and glucocorticoid deficiency-1, besides the more obvious links with adrenocortical carcinoma. Using the cytoview framework, ACTHr can be represented by 22.5.0.4.7.17.63, where 22 indicates it is a melanocyte, 5 indicates it is the cytoplasm, 0 indicates no specific sub-organelle defined at the moment, 4 stands for molecular networks (adenylate cyclase pathway for example), 7 indicates the macromolecular-metabolite complex responsible for the inherent biochemical reaction(s), 17 indicating the individual molecular constituents (proteins, genes, metabolites further defined through a sub-classification as 17_p, 17_g and 17_m keys) and 63 indicating the type of three dimensional structure(s) of the protein in the above layer. Mutations in ACTHr such as D107N (Clark *et al* 1997) [defined at the seventh field in the dotted representation, which leads to loss of binding of adenylate cyclase (one from the list of 17 m in the sixth field, resulting in loss of appropriate complex formation (fifth field) and hence dysfunctional molecular networks (fourth field), further linked to alteration in a feature of the cytoplasm (second field; abnormal cytoplasmic trafficking)], and further to the dimensions of the melanocyte itself (first field), which has been correlated with adenoma. For e.g. the size of a melanocyte grows several-fold in pituitary adenoma, in which there is also a concurrent increase in biochemical profiles of certain metabolites (Selvais *et al* 1998). This example illustrates how computational models can capture flow of information among different levels of hierarchy such as from particular mutations in ACTHr that may ultimately lead to skin cancer. Systematic navigation through the different layers enables us to relate changes at the genotypic levels to those at the phenotypic levels, thus making this framework useful in obtaining insights into physiology or disease processes.

Supplementary table. Overview of cell organelles

	Appearance	Size	Location	Remarks
Centrioles	Tubular bundle of microtubules	15 nm (dia) 350–500nm (long)	During cell division found near nucleus at approx. center of cell	Usually 2 centrioles are present in the centrosome
Cilia	Hair like motile organelle	10 μm (long) 0.25 μm (dia)	Locomotory organ present on periphery of membrane	
Flagella	Helical Rigid structure with rotatory motor at base	200 μm (long) 0.25 μm (dia)	Locomotory organ present on membrane	
Endoplasmic Reticulum	Interconnected and ramifying tubules (Smooth), Sacs (Rough)	Smooth-150 nm (dia) Rough–100nm in depth	Throughout cytoplasm	Rough ER associated with ribosomes
Golgi Apparatus	Stack of discoid saccules but may form complex networks.	Variable but approx. 2500 nm	Near Nucleus and Centriole	Usually surrounded by vesicles
Intermediate Filaments	Rod shaped	10–12 nm (dia)	Throughout Cytoplasm	Solid, Smooth surfaced and unbranched filaments
Lysosome	Approx. spherical with single membrane	200–500 nm (dia)	Usually towards periphery of cell near golgi complex	Contains powerful digestive enzymes
Microfilaments		25 nm (dia), 5 nm (thick)	Throughout Cytoplasm	
Mitochondria	Spherical, ovoid, filamentous structure (double layered)	2–6 μm (length), 0.2 μm (dia)	Scattered throughout Cytoplasm	Inner membrane in folds to create bulkhead-like structures called cristae
Nucleus	Spherical or ovoid may be lobed (double layered)	5–10 μm	Approx. Center of cell(animal cell) shifted towards periphery (plant cell)	40–70 nm intermediate space nuclear pore 70 nm (dia)
Ribosome	Approx. spherical bodies often arranged in strings or spirals	25 nm (dia)	Outer surface of Rough Endoplasmic Reticulum	Associated with mRNA, involved in protein synthesis

Acknowledgements

We are grateful to Prof. Vijayan for constant encouragement, support and useful discussions. Financial support from Department of Biotechnology (DBT), New Delhi is gratefully acknowledged. Use of facilities at the Super Computer Education and Research Centre, Bioinformatics Centre and Interactive Graphics facility supported by DBT is also acknowledged.

References

Aleshin A E, Zeng C, Bourenkov G P, Bartunik H D, Fromm H J and Honzatko R B 1998 The mechanism of regulation of hexokinase: new insights from the crystal structure of recombinant human brain hexokinase complexed with glucose and glucose-6-phosphate; *Structure* **6** 39–50

Ashburner M, Ball C A, Blake J A, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K *et al* 2000 Gene ontology: tool for the unification of biology. The Gene Ontology Consortium; *Nat. Genet.* **25** 25–29 (*http://www.geneontology.org/*)

Bard J, Rhee S Y and Ashburner M 2005 An Ontology for cell types; *Genome Biol.* **6** R21

Beuschlein F, Fassnacht M, Klink A, Allolio B and Reincke M. 2001 ACTH- receptor expression, regulation and role in adrenocortical tumor formation; *Eur. J. Endocrinol.* **144** 199–206

Bourdeau I and Stratakis C A 2002 Cyclic AMP-dependent signaling aberrations in macronodular adrenal disease; *Ann. N.Y. Acad. Sci.* **968** 240–255

Clark A J L, Cammas F M, Watt A, Kapas S and Weber A 1997 Familial glucocorticoid deficiency, one syndrome but more than one gene; *J. Mol. Med.* **75** 394–399

Cormen T H, Leiserson C E, Rivest R L and Stein C 2001 *Introduction to Algorithms* Second edition (MIT Press)

Hunter P J and Borg T K 2003 Integration from protein to organs: The Physiome Project; *Nat. Rev. Mol. Cell Biol.* **4** 237–243 *(http://www.physiome.org.nz/)*

Hunter P J 2004 The IUPS Physiome Project: a framework for computational physiology; *Prog. Biophys. Mol. Biol.* **85** 551–569

Ideker T, Galitski T and Hood L 2001 A new approach to decoding life: systems biology; *Annu. Rev. Genomics Hum. Genet.* **2** 343–372

Kiehl T R, Mattheyses R M and Simmons M K 2004 Hybrid simulation of cellular behavior; *Bioinformatics,* **20** 316–322

Kremer J R, Mastronarde D N and McIntosh J R 1996 Computer visualization of three-dimensional image data using IMOD; *J. Struct. Biol.* **116** 71–76

Lin C J, Jorge A A L, Latronico A C, Marui S, Fragoso M C V, Martin R M, Carvalho F M, Arnhold I J P, Mendonca B B 2000 Origin of ovarian steroid cell tumor causing isosexual pseudoprecocious puberty demonstrated by the expression of adrenal steroidogenic enzymes and adrenocorticotrophin receptor; *J. Clin. Endocrinol. Metab.* **85** 1211–1214

Lloyd C M, Halstead M D and Nielsen P F 2004 CellML: its future, present and past; *Prog. Biophys. Mol. Biol.* **85** 433–450 (*http://www.cellml.org/*)

Loew L M 2002 The virtual cell project.; *Novartis Found. Symp.* **247** 151–160 (Virtual Cell. *http://www.nrcam.uchc.edu*)

Mesiano S, Fujimoto V Y, Nelson L R, Lee J Y, Voytek C C and Jaffe R B 1996 Localization and regulation of corticotrophin receptor expression in the midgestation human fetal adrenal cortex: implications for in utero homeostasis; *J. Clin. Endocrinol. Metab.* **81** 340–345

Mountjoy K G, Robbins L S, Mortrud M T and Cone R D 1992 The cloning of a family of genes that encode the melanocortin receptors; *Science* **257** 1248–1251

Rosse C and Mejino J L Jr 2003 A reference ontology for biomedical informatics: the Foundational Model of Anatomy; *J. Biomed. Informat.* **36** 478–500 (*http://sig.biostr.washington.edu/projects/fm/AboutFM.html*)

Selvais P, Donckier J, Buysscheart M and Maiter D 1998 Cushing's disease: a comparison of pituitary corticotroph microadenomas and macroadenomas; *Eur. J. Endocrinol.* **138** 153–159

Solminski A, Erma G and Mihm M 2004 ACTH receptor, CYP11A1, CYP17 and CYP21A2 genes are expressed in skin; *J. Clin. Endocrinol. Metab.* **81** 2746–2749

Thomaseth K 2003 Multidisciplinary modelling of biomedical systems; *Comput. Methods Prog. Biomed.* **71** 189–201

Tomita M, Hashimoto K, Takahashi K, Shimizu T S, Matsuzaki Y, Miyoshi F, Saito K, Tanida S, Yugi K, Venter J C and Hutchison C A 3rd 1999 E-CELL: software environment for whole-cell simulation; *Bioinformatics,* **15** 72-84. (*http://www.e-cell.org/*)

Zwermann O, Beuschlein F, Klink A, Stahl M and Reincke M 2004 The role of the ACTH receptor in adrenal tumors: Identification of a novel microsatellite marker; *Horm. Metab. Res.* **36** 406–410

*e*Publication: 6 July 2007