Before we can integrate multi-scale models, or indeed before we can adequately describe a single-scale model, we must have an agreed upon framework with which to describe the underlying biology. Without a unified biological description technology, the goal of multi-scale modeling, model (and module) sharing and reuse will be greatly inhibited.

Prof. Bassingthwiaghte's thoughtful first draft of an MSM white paper covers, primarily, the later stages of model definition, i.e., the description of modular components largely defined at the computational level. It seems to me that what is first needed is a definition of the modular components at the biological level. This "Biological Description" is agnostic to the details of any particular computational framework. The Biological Description is both the first level of model description and is also the first level of model validation. The Biological Description should be in the language of modern biology and may contain little, or even no, mathematical details, let alone computational instantiation details.

The Biological Model description level is the natural level for model sharing (and reuse) and linking of models across multiple scales. Since the Biological Model is agnostic to computational details the details of implementation of a particular module are irrelevant to the task of combining (or reusing) models.

I believe that this Biological Model, in a form that a wet-lab biologist would be comfortable with, is a required precondition to developing interoperable multiscale models.

Computational biologists already often use an *ad hoc* ontological description of their models. Computational biology publications often include a table similar to Fig. 1 (Shirinifard A, Gens JS, Zaitlen BL, Popławski NJ, Swat M, et al. (2009) 3D Multi-Cell Simulation of Tumor Growth and Angiogenesis. PLoS ONE 4(10): e7190). The table in Fig. 1 represents an "ontological description" of the particular model but it was created without an actual ontology (since a suitable ontology does not exist). The *ad hoc* nature of this common publication technique makes locating, defining, validating and sharing the model extremely difficult.

Figure 1: Representative *ad hoc* ontological description of a computational biology model.

Cells	Behaviors
Tumor cells	
Normal	-proliferate
	-consume oxygen
	-change to hypoxic
	-change to necrotic
Hypoxic	-proliferate
	-consume oxygen field
	-change to normal
	-change to necrotic
	-secrete long-diffusing proangiogenic field $V(\vec{i})$
Necrotic	-shrink
	-disappear
Endothelial cells	
Vascular	-consume oxygen field
	-supply oxygen field at partial pressure $P_{blood}^{vascular}$
	-secrete short-diffusing chemoattractant field $C(\vec{i})$
	-chemotax up gradients of field $C(\vec{i})$
	-elastically connect to neighboring
	vascular and inactive neovascular cells
	-lose elastic connections, when $l > l_{max}$

The *ad hoc* nature of the "ontological description" that most computational biology papers use also tends to obscure the choices that were made in translating the Biological Model into a computational model. Details, such as the representation of space and time in the computational instantiation, are not included in the "ontological description" but instead must be culled from the text of the paper or worse yet, from references within references in the paper.

I believe therefore that before any significant progress can be made in multi-scale modeling and model (module) sharing, reuse and validation there must first be a useable formal ontology that can, at least,

describe the observed biology and the types of models that a wet-lab biologist is familiar with. This formal ontology should, at some level, be completely agnostic to both the mathematical description of the system and to the computational framework(s) in which it might be instantiated.

There are many, many kinds of "models"

At the risk of overstatement I think it is safe to say that all of science is "models". In the biological domain, models range from simplified "blob" diagrams that might be found in a freshman biology textbook, to complex interaction maps such as a KEGG pathway, to even more complex representations such as the electro physiochemical models of the heart. Importantly, biological experiments and experimental results are also models. Indeed it is the wet-lab "models" upon which most computational models are not only based upon but also validated against.

In order to adequately describe a computational biology model it must be possible to describe it first in purely biological terms. At the biological level the model (or module) is guaranteed to be "mineable", shareable and "hot-swappable" with other models (or modules). In addition, it is the biological description that defines how models at different spatial (or temporal) scales interact.

Ontologies for Biology

A plethora of biological ontologies have already been developed. These ontologies range from exhaustive "naming authority" type ontologies such as FMA and GO to "toolkit" ontologies like SBML and OPB. The major difference between a "naming authority" and "toolkit" ontology is their ability to be used to *instantiate* a description of a particular biological system. Ontologies like FMA and GO provide a controlled vocabulary of concepts and generally place those concepts in some type of hierarchical tree. For example, FMA can be used to provide the name "hepatocyte" to the major cell type of the liver. In addition, the FMA trees locate that cell ("is part of" the "portal lobule" which "is part of" the "liver") and may also provide cell lineage or other information. However, FMA does not provide terms needed to specify that a model might consist of multiple hepatocytes and those cells have a characteristic geometric distribution. In "naming authority" ontologies a new concept, for example a cell type in FMA or gene in GO, is generally added to the ontology and becomes a permanent component of that ontology.

"Toolkit" type ontologies are designed to provide a set of basic object types and a set of possible relations between those types. For example, to instantiate a hepatocyte in OPB one might describe the size and shape of the cell, its neighbors and the types of processes it can undergo. FMA does not already have a concept of "hepatocyte", indeed it really doesn't contain the concept of a cell. Once a cell is described using the terms of OPB the resulting description does not become a permanent part of the ontology.

I believe that MSM (and computational biology modeling in general) requires the creation of a suitable "mash-up" ontology that can be used to describe Biological Models and computation models.

Ontology development is a time consuming task. To develop the complete ontology needed for describing both wet-lab experiments and results and computational models is a daunting task. It may be possible to largely avoid the creation of a large and robust ontology by instead creating, essentially, a

markup language and a fairly small ontology. The majority of the ontological terms (and relations) can be "inherited" from existing ontologies. For example, a model (wet-lab or computational) should use the excepted names for tissues, cells, genes and proteins. In a markup-up type description those terms can be extracted (linked to) existing "naming authority" ontologies such as FMA and GO. This implementation has several important advantages. 1) It automatically creates "crawlable" linkages to rich data repositories (not only the referent ontology but to linkages contained therein). 2) It insures that named entities in models use accepted names, which is a requirement for efficient mining of the models. 3) It significantly reduces the work required to develop a useable "model description" ontology. 4) It insures that the basic structure and concepts are "biologically" defined (as opposed to computationally defined).

For concepts relating to spatial, temporal and energy (K, k, ΔG …) terms OPB might be a suitable external ontology that can provide the needed terms and relationships.

The required depth of the "ontology" or mark-up language is unclear to me. It must be able to at least specify observable biology. This would include the ability to describe both objects (small molecules, cells, proteins, organs etc.) but also biological processes like transport, mobility, binding, development, differentiation, growth, death, signal generation and receiving and so forth. It seems likely that the language will also need to be able to describe the fundamental mathematics of a particular process. For example, kinetic and diffusion processes defined mathematically (nut in a computationally agnostic way).

As the need to describe the model progresses from the Biological Model to the computational instantiation of the model the types of relationships may explode in number. It seems reasonable though that certain characteristics of the computational instantiation should be describable. For example, how spatiality is treated. Is space discretized or continuous? Does the model even have a concept of space? The same for time, are there time steps (if so how big) or not? How exactly are fundamental biological processes instantiated? If the model includes cells that grow what controls the growth and what formula is used to calculate the growth as a function of time?

Even if it is found that an ontological description of all of the computational details is not practical it is still critically important that the Biological Model is described using the structured language.

Some work has already been done on "mash-up" type ontologies for computational biology. SemSim and SemGen (Univ. Washington) include many of the ideas I have presented here. Extending those tools to describe wet-lab assays and results and enhancing their ability to describe Biological Models would leverage existing software and may provide a rapid route to a functioning ontology and markup language tool. Adding the ability to publish the models via the SW may provide a rapid route to model dissemination, validation and reuse.

Advantages of an "Ontological" description at the Biological Model and other levels

If models (biological, computational …) can be described in a structured way then the models can be published using semantic web (SW) technologies[1]. This method of publication makes the models searchable without a user needing to know where to actually look for the models. The model might be in a well maintained computational model repository (like the BioModels Database), or it might be on a server maintained by an individual research lab. The data might be a Biological Model, or a computational model, or an entry in a vast repository of biological assay results such as the EPA's ToxCast system. (Which should be describable using our imagined ontology and publishable via the SW, instead of being buried in a web accessible database which cannot be automatically mined.)

An important benefit of our imagined ontology is that experimental data is describable, SW publishable and searchable (mineable). Since that data is often used to parameterize models finding it should become easier.

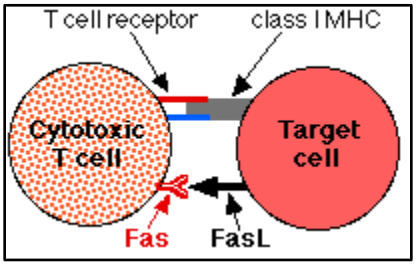
Defining the Ontology (or markup language)

The ontology perhaps should be developed in two arms. The first arm would be the selection (possibly with certain restrictions) of the reference ontologies. The second arm is the identification of key terms and relationship that are either absent in the reference ontologies or that are of such fundamental import to experimental and computational biology that a redefinition is warranted.

The key terms and relationships could perhaps be defined using two pathways. One pathway would start with wet-biology and biologist. What are the fundamental objects and processes that they think must be included? The second pathway would start with existing computational biology models and work backwards. It is not necessary that the ontology can describe all the details of the computational model but it appears to me that most computational biology groups have already (if unconsciously) followed the first pathway and there modules and code have a tendency to mirror basic biological concepts. Working backwards from exiting computational models leverages that knowledge.

A Really Simple Sample Model and Markup

What exactly would the process look like and what would it create? We will start with the very simple model, like what might be found in a basic biology text book. Shown below is the interaction of a Cytotoxic (killer) T-cell with an infected host cell (http://users.rcn.com/jkimball.ma.ultranet/-BiologyPages/T/Transplants.html).



[1] Yes, the semantic web is largely vaporware, or at least it is as a general method of coding information on web pages. However, in this case only the pages that our community generates need to conform to the standard.

We start with describing the objects in the model, two cell types and four membrane bound proteins. We can use FMA to describe (unambiguously name) the cells and GO to describe (unambiguously name) the four proteins. We can use terms and relationship from OPB to describe the binding events. In pseudo XML like format[2];

```
<XML=”MSM pseudo Code”>
<External “MSM” link to MSM>  <!—our imaginary ontology -->
<External “FMA” link to FMA>
<External “CL”  link to CL >
<External “GO”  link to GO >
<External “OPB” link to OPB>
<Cell type=FMA:"PREFERRED NAME=T-cytotoxic cell" FMA:"FMAID=70573">
   <protein location=FMA:"cell membrane" name=GO:"T cell receptor">
   <protein location=FMA:"cell membrane" name=GO:"Fas">
</Cell>
<Cell type= FMA:"PREFERRED NAME=cell" FMA:"FMAID=68646">
   <protein location=FMA:"Plasma membrane" name=GO:"Class I MHC">
   <protein location=FMA:"Plasma membrane" name=GO:"FasL">
</Cell>
<Interaction type=binding>
   <entity1 name=GO:"T cell receptor">
   <entity2 name=GO:"Class I MHC">
   <FMA:Dissociation Constant="1e-6” Units=”Molar”>
</Interaction>
<Interaction type=binding>
   <entity1 name=GO:"Fas">
   <entity2 name=GO:"FasL">
   <FMA:Dissociation Constant="1e-7” Units=”Molar”>
</Interaction>
```

The markup above adequately recapitulates the details of this very simple Biological Model. Though simple the model still contains a significant amount of information (“knowledge”). To extend this markup to a particular computational framework might involve including terms to describe how the binding is calculated; mass action law, stochastic etc. A particular computational framework might also need to describe the movement of the cells and that requires some definition of space, distance and time. However, regardless of the markup that is added to describe the computational instantiation of the model the Biological Model, and its description, is constant.

In the above description we have left out the several details of this Biological Model. For example, 1) binding of Fas to FasL trigger apoptosis of the Target (infected) Cell. 2) The T-Cell Receptor (TCR) represents a complex entity that is not describable in the way most genes and gene products are described, therefore the “naming authority” link to FMA does not have the same information content as most gene/gene product names. The mature form of the gene is the result of an irreversible genomic rearrangement that occurs as the T-Cell matures. 3) The TCR – Class I MHC interaction is actually mediated by a peptide fragment present in the Target Cell. 4) “Class I MHC” represents a heterodimeric protein complex.

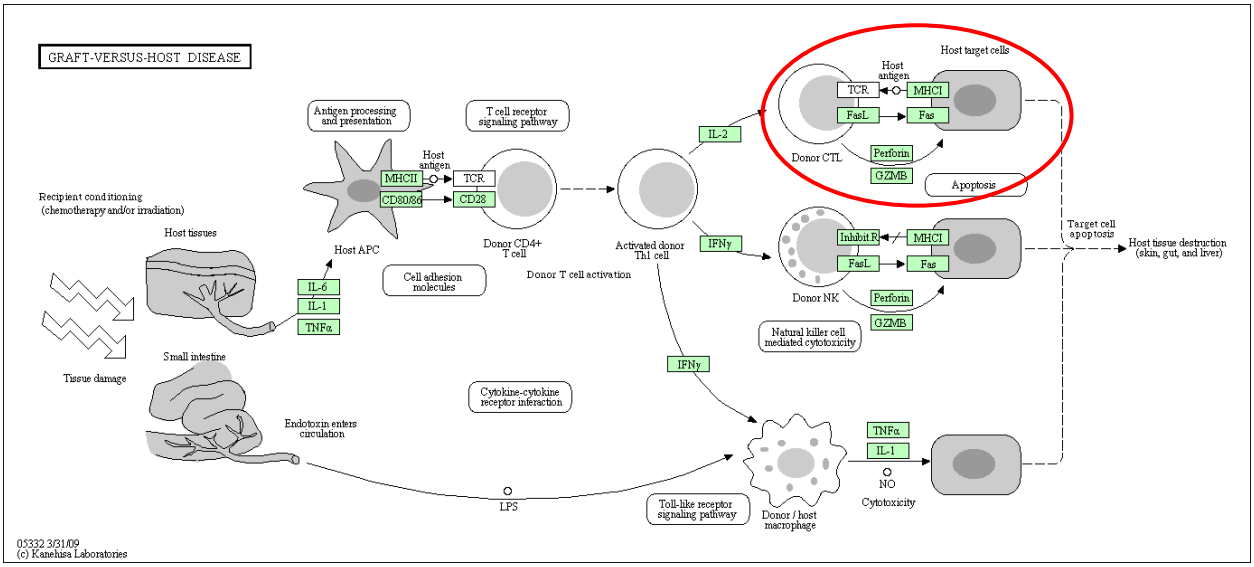
Besides describing the basic model the markup above provides “crawlable” linkages to other information. For example, following the FMA linkage for FMA:"PREFERRED NAME=T-cytotoxic cell" provides the synonyms for this cell type: “Cytotoxic T-lymphocyte”, “Cytotoxic T cell”, “Killer T lymphocyte”, “Killer T cell” and “Cytotoxic T lymphocyte”.

[2] The style should probably by OWL/RDF instead of XML.

The sample markup above gives a very simple pseudo-description of the binding interaction. The ontology should also allow more detailed models to be incorporated into the model's description. As long as there is a systematic way to describe the components, preferably via linkages to existing ontologies, then any markup can be included in the model. For example, an SBML type model could be included.

A More Complex Sample Model and Markup

The KEGG pathway (http://www.genome.jp/kegg-bin/show_pathway?map05332) below is an example of a more complex Biological Model.



In this model there are several cell types and about a dozen proteins. This model contains as a sub-model the very simple model we examined earlier; the recognition of an infected cell, via the T-Cell Receptor to Class I MHC interaction, by a cytotoxic T-Cell (circled in red). Note that the names used in this KEGG figure do not exactly match the names used in the first model. To make a model sharable it is critically important that this type of discrepancy is avoided. The use of reference ontologies when the model is described in our markup/ontological description regularizes the names and makes it possible to identify the first model as a sub-model of the second. In other words, consistent naming within the model description makes it possible to find candidate existing models that can be reused.

The variability in the naming of cells, tissues, proteins and genes is a significant barrier to identifying and reusing models.

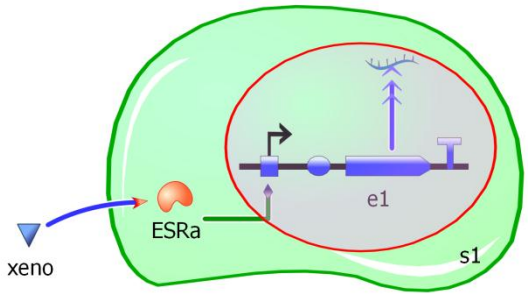
A Completely Different Type of Sample Model

It should be possible to encode simple Biological Models that are run as screens, including cases where the screens are high-throughput. For example, an Attagene Inc. assay being run for the EPA measures the transcriptional activation effects of xenobiotics binding to the Estrogen Receptor Alpha. A very small fragment of the data table for the Estrogen Receptor Alpha assay is shown below.

Data Set Name: Attagene
Column 1: Unique chemical identifier (320 unique values)
Column 2: CAS Registry Number (309 unique values)
Column 3: Chemical name (309 unique values)
Columns 4-N: Assay data. For a description of the assays, see the Assay Definitio
Values are LEL (lowest effective level)
Values are in microM
Inactive chemical-assay combinations are indicated by a value of "-"

SOURCE_NAME_SD	CASRN	NAME	Ahr_CIS	AP_1_CIS	BRE_CIS
DSSTOX_40310	136-45-8	2,5-Pyridinedicarbox	-	-	-
DSSTOX_40542	90-43-7	2-Phenylphenol	-	-	58
DSSTOX_40375	55406-53-6	3-Iodo-2-propynylbut	-	-	-
DSSTOX_40294	135158-54-2	Acibenzolar-S-Methyl	38	-	-
DSSTOX_40338	50594-66-6	Acifluorfen	-	-	-
DSSTOX_40339	15972-60-8	Alachlor	-	7.4	12
DSSTOX_40344	33089-61-1	Amitraz	-	-	-
DSSTOX_40299	101-05-3	Anilazine	62	45	-
DSSTOX_40347	86-50-0	Azinphos-methyl	-	44	23
DSSTOX_40348	131860-33-8	Azoxystrobin	-	3.8	37

This is a cell based assay and the compounds tested must cross, at least, the cell membrane. The Biological Model might be represented as;



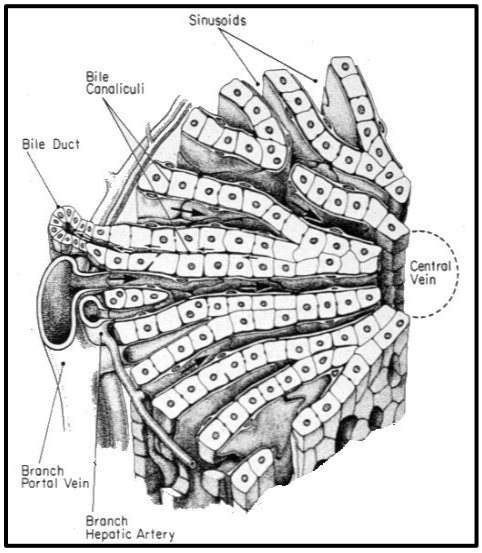
"Xeno" represents a single compound from the assay table. The molecular characteristics of "Xeno" can be obtained from other tables associated with the EPA ToxCast system. (ToxCast is not SW compatible but hopefully will be eventually.)

```
<XML="MSM pseudo Code">
<External "MSM" link to MSM>  <!—our imaginary ontology -->
<External TCC link to ToxCastCompounds >
<External TCA link to ToxCastAssays >
<External "GO"  link to GO >
<Cell type=FMA:"PREFERRED NAME=cell" FMA:"FMAID=68646">
   <protein location=FMA:"cytosol" name=GO:"ESRA">
</Cell>
<Interaction type=binding>
   <entity1 name=GO:"ESRA">
   <entity2 name=TCC:"Name=Acibenzolar-S-Methyl" id=TCC:"CASRN=135158-54-2">
   <createdEntity>entity1:name+entity2:name</createdEntity>
   <MSM:LEL="38" Units="microMolar">
</Interaction>
<transport>
   <entity name=TCC:"Name=Acibenzolar-S-Methyl" id=TCC:"CASRN=135158-54-2">
   <type = passive>
   <cross= type=FMA:"PREFERRED NAME=cell" FMA:"FMAID=68646" FMA:"cell membrane">
</transport>
<transport>
   <entity name=entity1:name+entity2:name>
   <type = active>
   … describe translocation to the nucleous of the complx
</transport>
<Interaction type=binding>
   <entity1 name=MSM: entity1:name+entity2:name >
   <entity2 name=GO:"Gal4" type=GO:CDS>
```

```
</Interaction>
```

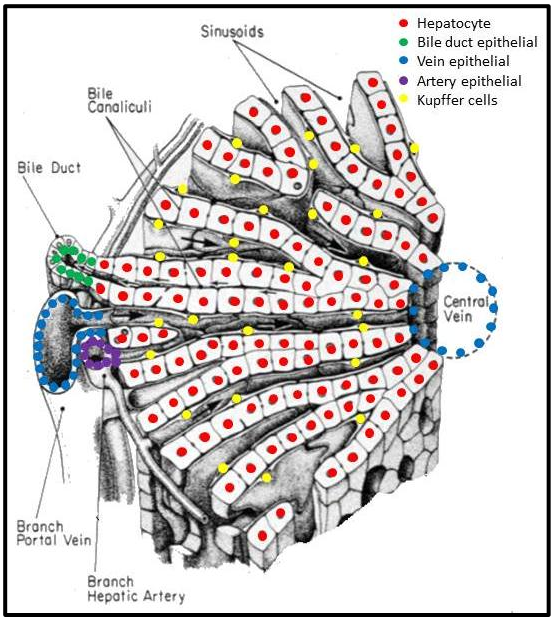
Once the Biological Model is described it can easily be changed to represent other rows in the assay results table by simply changing the compound reference and the results value.

Another Completely Different Type of Sample Model



Using our imagined ontology and mark-up language it should be possible to describe tissues and organs. At right is a drawing of a liver lobule. In this drawing there are two main cell types shown; Hepatocytes and Kupffer cells (liver macrophages). In addition the central and peripheral blood vessels are shown diagrammatically. In the case of the liver, the arrangement of the cells forms the sinusoids through which blood flows from the periphery to the central vein of the lobule. Blood flow through the sinusoids mixes venous blood from the GI tract with oxygenated blood from the hepatic artery. A Biological Model of the liver might include both the cell types and their spatial distribution along with a description of the blood flow pattern.

As with the previous examples, development of the Biological Model should start with a description of the cell types involved (with correct linkages to a naming authority such as FMA or CL) along with spatial descriptors defining each cell. In addition, a description of the blood flow (with "blood" also properly identified using, for example, FMA) would also be of use. The required level of detail describing the blood flow might range from simply indicating that the periphery is at higher pressure than the central vein, or the description might include terms such as viscosity and the cross-sectional area of the sinusoids (described using terms from OPB).



This particular model also introduces the need to be able to describe adhesion between cells. In the case of the liver, cell-cell adhesion is responsible for maintaining the basic topology. Adhesion could be described in the Biological Model as simple "stickiness" between the pair of cells. Or, if there is sufficient knowledge about the interaction the adhesion might be specified using explicit adhesion molecules (similar to our first model).

Carrot vs. Stick & Nasty Realities

- Getting people to accept an ontology, or markup language, is a challenge. A tool that provides both easy markup as well as other immediate tangible results will greatly increase the chances that the markup and ontology are widely adopted.
- It is likely that as some level of detail it will get to be extremely difficult to completely describe a computational model.

Stretch Goals

- Mining of masked models (e.g., remove cell and protein names); connectivity as a generic model or meta-model. Similar to mining SBML/JARNAC models using the network connectivity matrix. In our first simple sample model if the cell and protein names are removed we are left with a generic model of cell-cell adhesion mediated by two pairs of adhesion molecules.

James Sluka
Biocomplexity Institute
Indiana University
Bloomington, IN