ICBO: International Conference on Biomedical Ontology
July 28-30, 2011 · Buffalo, NY, USA

Composite Annotation for Heart Development

Tariq Abdulla[1], Ryan Imms[1], Jean-Marc Schleich[2], Ron Summers[1]

[1]Research School of Systems Engineering, Loughborough University, Loughborough, UK
[2]LTSI Signal and Image Processing Laboratory, Université de Rennes 1, Rennes France

Abstract. This paper describes progress made in combining multiple ontologies in a post-coordinated approach. This is applied to the annotation of phenotypes relevant to heart development and congenital heart diseases. The aim is to provide coordination for multiscale modeling and simulation that is presently being conducting in this field. Cardiac development is well understood within discrete levels of analysis. The application of the multiscale framework gives added value by unlocking relationships between genetic-based information at one level of analysis and the emergent phenotype at the cell and organ levels of abstraction. The challenge for a semantic representation is that this field encompasses multiple spatial and temporal scales. As a consequence, relevant terms come from a wide range of biomedical domains, and are therefore contained in several reference ontologies. The strategy for composite annotation provides a method for linking between multiscale measurement and modeling.

Keywords: Cardiac, Embryo, Morphogenesis, Imaging, Congenital, Multiscale.

1 Introduction

The explosion of data generated by many different fields of biomedical research has led to an increased focus on multiscale modeling and simulation, which provides the abstraction necessary for representing biological systems in a tractable way. The growth of computational biology is such that modeling and simulation now themselves represent a challenge in integration. The Physiome and VPH research community collaborate to provide curated model repositories of biochemical reaction networks (in SBML [1]) and biophysical mechanisms (in CellML [2]). These then have the potential to be reused in whole or part by other modelers working on different problems, potentially on different platforms, in different parts of the world. The open modeling paradigm is now so influential that some publishers advise depositing a model in these repositories as part of the publication process [3]. However, the reuse of such models will be greatly facilitated if they are represented with a common semantics, so that it is clear how the entities and parameters in one model relate to those in other models. If the same system of semantics is also used for representing experimental results, model validation will also be greatly facilitated.

In parallel to the rise of multiscale computational modeling, a suite of reference ontologies are being developed under the umbrella of the OBO Foundry [4], which provide increasingly good coverage of biomedical concepts at different levels of spatial and temporal scale. Initially, this grew from the coordination of heterogeneous databases that record the characteristics of gene products, primarily with the Gene Ontology (GO). Reference ontologies are now used for annotating a wide variety of biomedical knowledge sources. These sources include images, database entries, publications, computational models and simulation results. By keeping reference ontologies well-bounded and essentially orthogonal the OBO Foundry minimizes logical inconsistencies and confusion over which ontology to use.

For many applications, there is a need to combine terms from multiple reference ontologies, in order to create a composite term suitable for a particular annotation. This can either be done by defining terms in application ontologies as equivalent to a composition of reference ontology terms (pre-composition); or through post-composition, whereby the annotator can compose terms on the fly, and add them to a repository of composite terms. While the former approach is less complex for the annotator, the latter approach is more flexible.

Multiscale modeling efforts have focused mainly on the physiology of adult organ systems. Post-composed annotation of models has so far been applied only to physiological models with fairly simple physical properties [5]. The work reported here aims to tailor the multiscale framework for application to morphogenesis of the human embryonic heart. The levels of temporal and spatial scale applicable to heart development, and methods of representation are illustrated in Fig. 1. Computational modeling approaches that can be applied at different levels of scale are shown, as well as markup languages that enable a degree of model sharing between different platforms. The XML languages force a declarative expression of the components of a model, which allow it to be interpreted by different platforms. It is straightforward to annotate XML, and create an explicit link between entities in the model and external identifiers, that can be interpreted by software agents. In contrast, procedural code might be annotated with in-line comments that need a human reader to interpret them.

Along the bottom of Fig. 1, the ontologies applicable to different levels of scale are illustrated, which can be used for annotation of different model components. These ontologies are split between occurents, independent continuants and dependent continuants, following BFO and OBO Foundry conventions. By making this high level distinction, the OBO community has created a clearly defined boundary between the spatial and temporal domains. As simulation models comprise both domains, it is necessary to either combine terms in a post-composition approach, or make use of an application ontology for the annotation of a particular type of model.

The example process illustrated in Fig. 1 is epithelial to mesenchymal transition (EMT) in endocardial cushion growth. A signaling pathway within a single cell might be represented as an ODE within SBML. The interactions of cells and their chemical signaling might be represented with PDEs or stochastic Petri nets. A simulation of a larger numbers of cells is likely to use some form of agent based modeling. Finally, at the level of the developing heart tube as an anatomical component, finite element and multiphysics simulation may be used, to understand the relationships between mechanical properties of the heart walls (affected by EMT), its function as a pump and its looping morphology. The EMT process, and its central importance to heart development, is described in Section 2 and is used in the examples of composite annotation in Sections 3 and 4.

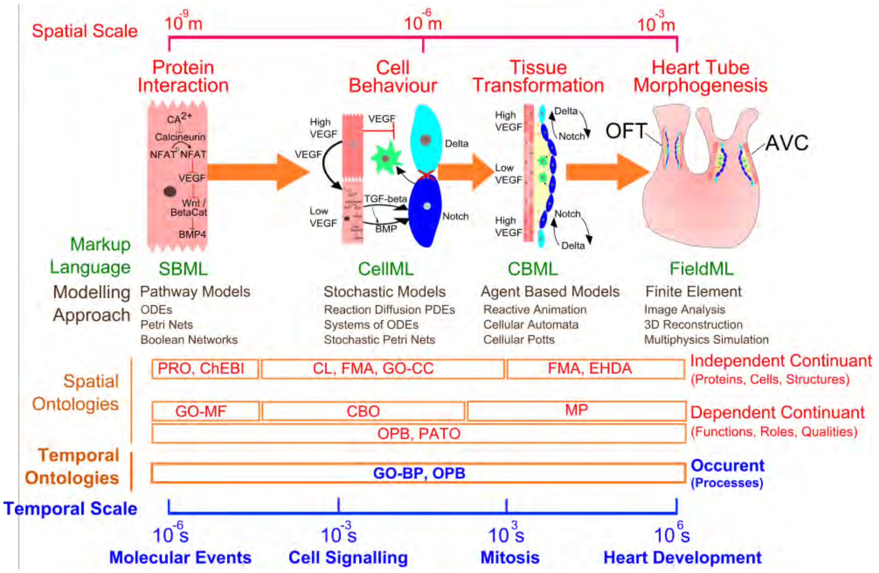

Figure 1. Spatial and temporal scales of heart morphogenesis modeling. The modeling framework encompasses spatial scales from 10^{-9}m (proteins) to 10^{-3}m (the primitive heart tube), and temporal scales from 10^{-6}s (molecular events) to 10^{6}s (weeks of heart development). All acronyms are defined in the Appendix.

2 Heart Development

2.1 The Anatomic Level

The development of the embryonic heart commences in week 2 of gestation and is fully formed by week 8. This process is well documented [e.g. 6]. Week 2 of foetal life provides the first milestone of cardiac development when the two endocardial tubes that form the primitive heart fuse together. At this stage of development the first cardiac muscle contractions occur, giving rise to both blood circulation and electrophysiological signals that form a primitive electrocardiogram [7].

The embryonic heart tube is composed of an inner layer of endocardium, an outer layer of myocardium and a middle layer of extra-cellular matrix termed cardiac jelly (Fig. 2). In two restricted areas of the heart tube – the outflow tract (OFT) and atrioventricular canal (AVC) – endocardial cells adopt a mesenchymal phenotype and invade the cardiac jelly. These restricted swellings are termed 'endocardial cushions' and are precursors for the heart valves and membranous septa.

The endocardial cushions begin to grow at embryonic day 26 (E26) in humans [8]. At the same time, the heart tube begins looping in an S-shape to the right. Two synchronised processes important in the understanding of congenital heart diseases are looping and aortic wedging. Looping is completed by E28, and is the first manifestation of asymmetry in the embryo. This repositioning constitutes a crucial step towards the morphology of the heart because it brings the future heart chambers and their inflow and outflow tracts into their relative spatial positions. Aortic wedging occurs as a consequence of the rotation of the myocardial wall of the OFT, itself secondary to the re-modeling of the inner curvature of the heart. Fusion of the cushions occurs at E32. In the OFT, parietal and septal cushions fuse forming the conal septum, which divides the aorta from the pulmonary artery. The conal septum is helical due to the rotation of the OFT. Upper and lower AVC cushions form the atrioventricular septum, the mitral valve and the tricuspid valve.

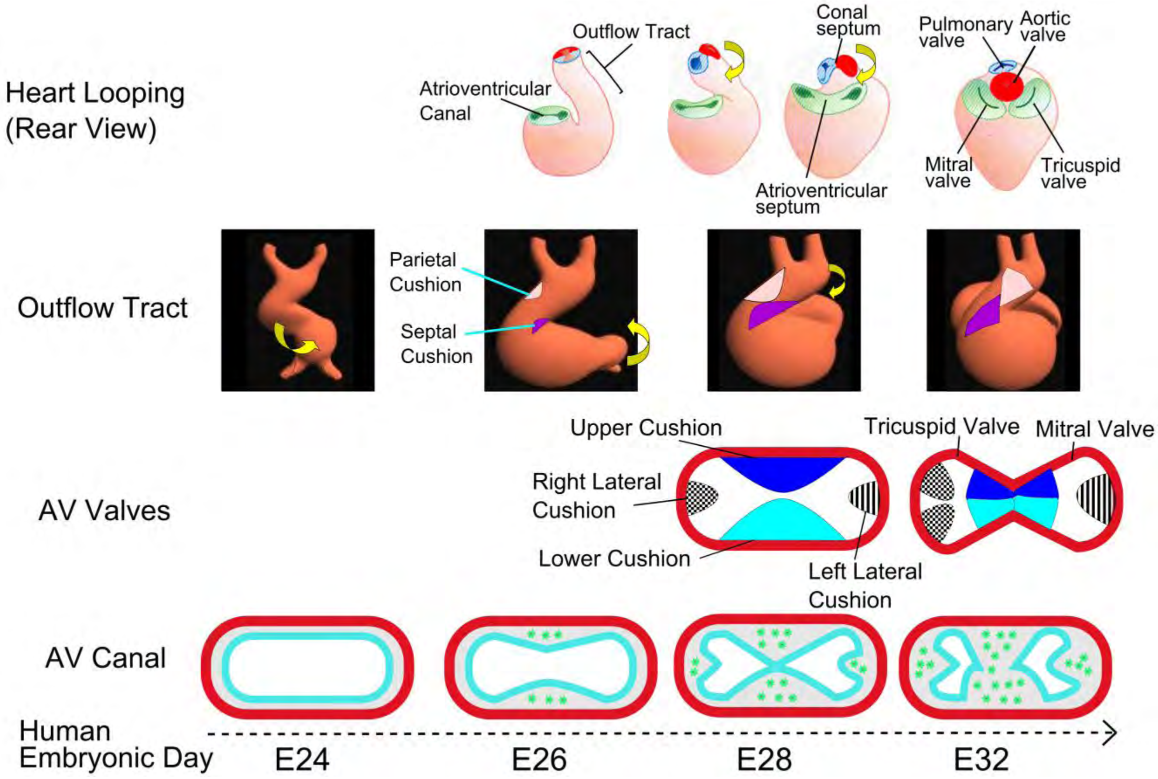

Figure 2. Detail of endocardial cushion growth and fusion. After [13].

2.2 The Cellular Level

The endocardial cushions grow by a process termed epithelial to mesenchymal transition (EMT). As the endocardial cushions play a role in forming much of the inner structure of the heart, it is apparent that abnormal EMT is a factor in many different types of congenital heart disease. These include valve, outflow tract and inter-ventricular septal defects (i.e. hole in the heart). During EMT, endothelial cells lose their adhesion to each other and invade the cardiac jelly, adopting a mesenchymal phenotype. This causes localized swelling on the inner surfaces of the embryonic heart. The study of EMT *in vitro* enables a controlled means for studying changes in cell properties, under the influence of different signaling proteins.

2.3 The Protein Level

Several signaling pathways have been identified as being important to heart development, in both the endocardium and the myocardium. Paracrine signaling acts over a short distance, perhaps between the two tissue types. These include the interactions of TGFß and BMP2 proteins, which are secreted by the myocardium in the cushion forming region. Principal among juxtacrine signaling pathways is the Notch pathway which controls pattern formation in many embryonic tissues, including those of the heart. In the endocardium, the Notch1 protein is expressed in the endocardial cushion forming regions, but not usually outside of those regions. In the myocardium the situation is reversed. Notch proteins have an additional role to play, because they activate the SNAIL family of proteins, which in turn inhibit transcription of the protein VE-Cadherin. As VE-Cadherin is one of the major proteins providing endocardial cohesion, activated Notch induces a loss of cohesion, which is part of the EMT process. It has been demonstrated *in vitro* that completion of EMT also requires BMP2 and TGFß proteins secreted by the myocardium [8]. Many types of congenital heart disease are associated with mutations in the Notch signaling pathway and this underlies the importance of Notch signaling to heart formation.

3 Methods

Developmental biology is a well established field of quantitative analysis. New results emerge every day from *in vitro* and *in vivo* high-throughput analysis, and add to the growing knowledgebase of genotype-phenotype associations. Heart morphogenesis is an area of particularly intensive research, as heart defects are among the most common type of congenital disorder. This has led to a recent expansion of the GO-BP to include a much broader range of biological process terms for heart development [9] and a corresponding initiative to increase the number of GO cardiovascular annotations. This represents a pre-composition approach, including creation of differentiation terms for 26 different cell types ('Endocardial cell differentiation', 'Pacemaker cell differentiation' etc.) Due to the logical structure of GO, these terms can be de-composed using cross-product extensions [10].

Post-composition has been applied successfully in annotating phenotypic descriptions. This makes use of a particular type of ontology composition: the Entity Quality (EQ) formalism. This extends entity terms from reference ontologies by describing them as the intersection of the entity with a relationship to a quality term in PATO. The entities are most often from species specific anatomy or developmental anatomy ontologies, but may also be a cell type from CL; a biological process, molecular function or cellular component from GO; or a molecular level entity from PRO or ChEBI. The EQ formalism has been used for investigating the evolution of phenotypic traits (phylogenetics) [11] and in integrating phenotypic annotations from multiple species [12], and in this way linking human diseases to mutant animal models [13]. In contrast, the Mammalian Phenotype (MP) ontology takes a pre-composition approach, which aims to include terms sufficient for phenotypic description within a single ontology [14]. This has been used successfully for the mouse and rat genome databases. The two approaches are not mutually exclusive, as MP terms could be defined as equivalent to EQ terms, when appropriate.

The post-composition approach has also

begun to be used for the annotation of biomedical simulation models. This is similar to the EQ formalism described above, but using the Ontology of Physics for Biology (OPB) rather than PATO. The OPB describes both physical properties and physical processes. This is because simulation models mainly represent the physical properties of biological entities. The SemGen tool enables modelers to annotate SBML or CellML code using OPB post-composition terms; although they must first be imported and compiled in the JSim modeling tool [15]. Once models are annotated in this way, a semantic comparison of several models can then be made through SemGen, automatically identifying entities that can be combined if models are merged. However, this approach to annotating models has only been applied to domains with well defined physical properties. It is not clear how well this would work for cell level modeling for example, where the physical properties that drive cell behavior are not fully understood.

It is straightforward to adapt the EQ formalism for developmental phenotypes. The initial step is to select the relevant ontologies for the domain, as well as the types of sources that might be annotated. The process for the domain of heart development is illustrated in Fig. 3.

PATO allows composite phenotype annotations such as 'endocardial cushion with decreased concentration of SNAIL protein', which are composed from the integration of multiple reference ontologies. OPB allows formalization of the physical properties of these composite annotations, such as the concentration of a particular protein in a particular endocardial cell, or the density of mesenchymal cells in an endocardial cushion. These terms can then be used to annotate variables in a computational model, or experimental data. PATO composites can also be mapped to disease classifications, such as OMIM.

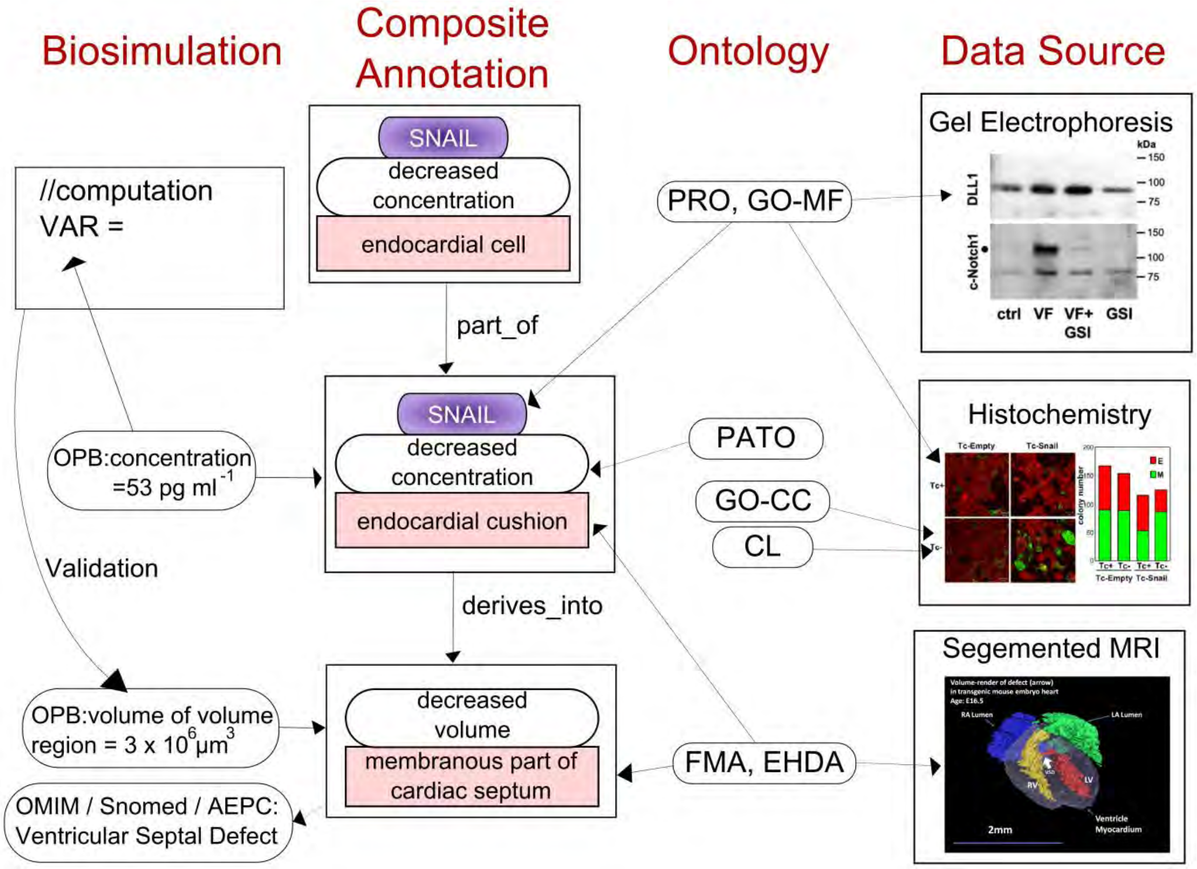

Figure 3. Schema for composite annotation. Refer to Appendix for acronyms.

OBO	OWL
intersection_of:	EquivalentTo:
PATO:0001163 ! decreased concentration	PATO:0001163
intersection_of: inheres_in	and (inheres_in some
PR:000015308 ! SNAI1	PR:000015308)
intersection_of: contained_in	and (contained_in some
CL:0002350 ! endocardial cell	CL:0002350)

Table 1. OBO and OWL representation of composite annotation.

4 Results

The topmost annotation shown in Fig. 3 can be represented in OBO or OWL format as detailed in Table 1.

Annotations will be represented using a simplified EQ syntax, using term labels rather than identification numbers. The annotation in Table 1 would be expressed as:

```
PATO:decreased concentration
<inheres_in> PR: SNAI1 <contained_in>
CL:endocardial cell
```

The OPB based composite annotation for the concentration parameter or measured value would be:

```
OPB:chemical concentration <property_of>
OPB:portion of molecules <composed_of>
PR:SNA1 <contained_in> CL:endocardial
cell
```

The same annotation could be used whether pointing to a model parameter, or an experimentally measured concentration. This suggests a method for leveraging the semantic relationships between very different types of information.

An EQ representation may be defined under a number of categories [11], with the examples below taken from the process of heart morphogenesis.

Monadic states are those that involve single entities or structures. For example, it has been previously shown that some congenital heart abnormalities are caused by an incorrect rotation of the outflow tract. This can be annotated in a general way as:

```
PATO:mislocalised_radially<inheres_in>
EHDA:outflow_tract
```

Relational states are those that describe a phenotype that exists between two entities or structures. The first example in this section was relational.

Quantitative states describe a measured value for a variable feature (e.g. size, area, count). For example, the volume of an endocardial cell would be annotated as:

```
OPB:volume region <inheres_in>
CL:endocardial cell <has_magnitude>
OPB:volume amount=3.2 <has_unit>
UO:microliter
```

5 Discussion

With post-composition, there is a lack of exact consistency in annotations between different annotators [12]. This is not always a major problem because, with sufficient guidelines, the differences are usually ones of specificity (e.g. did they use the FMA term ‘endothelium’, ‘endothelium of endocardium’ or ‘endothelium of aortic valve’?). These annotations are still valid semantically, but where a more coarse term is used there is a degree of information loss, to be avoided where possible. Restriction to terms of a specific domain and the use of customizable software tools for annotation improves consistency. An example of the latter is Phenote [11], an open source toolkit that facilitates annotation of biological data using OBO-format ontologies. However, it is still possible to have different perspectives on the same physiological phenomenon. For example, one decision might be whether the interest is in the decreased volume of the membranous septum, or the fact that the membranous septum is dysfunctional. From the perspective of exact volume quantification the actual size measurement is important, whereas in the more general disease classification the interest lies only in the fact that there is a dysfunction. There are often pre-composed terms in existing ontologies, which could also be made by post-composing terms from multiple ontologies. For example, in the MP ontology the term ‘abnormal outflow tract development’, could be composed as:

```
PATO:abnormal <inheres_in>
GO:outflow_tract_morphogenesis
```

The degree of variability possible is a key advantage of post-composition: congenital heart diseases are a spectrum of overlapping phenotypes, and it is necessary to have flexibility in the way they are annotated. This accuracy in genotype-phenotype annotation, while arguably more complex, is more beneficial to wider biological research than mere coding of defects for the sake of classification. However, the strategies are not mutually exclusive. An intriguing possibility is to map anatomical measurements (such as those determined from the MRI of congenital heart disease specimens) to disease classifications.

The challenge of reasoning over multiple ontologies remains a considerable one. Nonetheless, it is much more feasible to achieve data integration in this way than in any existing alternative. In particular if new ontologies were constructed for each application, with no semantic links to existing reference ontologies, then integrating across applications would be almost as cumbersome as not using ontologies at all.

6 Conclusion

It has been argued that simple annotations (e.g. a pointer to a single reference ontology class) are insufficient for annotating the variety of data sources that need to be integrated within the current multiscale modeling projects [16]. The variety of possible classes increases due to the need for more highly specified annotations. To this extent the post-composition approach is necessary for fully integrated multiscale annotation.

Multiscale modeling research has hitherto focused almost exclusively on adult physiology, with little attention to embryonic development, although this has begun to change. The team is in the process of developing multiscale simulations of endocardial cushion growth via EMT. Cell-tissue level modeling of EMT is achieved with Compucell3D, while signaling pathways are modeled using SBML. Combining knowledge gained from the information models (EQ formalism) allows the closure of the loop between physical experiments (real world) and computer based simulations (model world). As the EQ annotations of the model world map to their isomorphic physical counterparts in the real world it is possible to be unambiguous about referring to (say) endocardial cells or increased concentration of a given protein.

Creating accurate phenotypic descriptions, which retain their semantic context, and linking these to physical and biophysical measurements, provides a powerful means to assimilate information from a wide variety of sources and scales. To this end the team has access to a unique physical resource – over 50 post mortem heart specimens that have been diagnosed as tetralogy of Fallot. The intention of future work is to provide MRI data of these specimens to link between the primary evidence and degree of outflow tract rotation.

A further limitation to overcome is the lack of exact consistency in ontological annotations. Nevertheless, data sources of different types, at different scales have been identified, alongside the ontologies suitable for annotation, modeling methods at different levels, and initial guidelines for composite annotation. This demonstrates a method for creating a link between multiscale measurement and multiscale modeling that assists in closing the loop between physiological and genetic understanding of cardiac development.

References

1. Novere, N.L., Courtot M., Laibe, C.: Adding Semantics in Kinetics Models of Biochemical Pathways. In: 2nd International ESCEC Symposium on Experimental Standard Conditions on Enzyme Characterizations, pp. 137–154. Beilstein-Institut, Rhein (2007)
2. Beard, D.A., Britten, R., Cooling, M.T. et al.: CellML Metadata Standards, Associated Tools and Repositories. Phil. Trans. R. Soc. A. 367, pp. 1845–1867 (2009)
3. Courtot, L.C., Novère N.L., Laibe C.: BioModels.net Web Services, a Free and Integrated Toolkit for Computational Modeling Software. Briefings in Bioinformatics. 2, pp. 270–277 (2010)
4. Smith, B., Ashburner, M., Rosse, C. et al.: The OBO Foundry: Co-ordinated Evolution of Ontologies to Support Biomedical Data Integration. Nature Biotechnology, 25, pp. 1251–1255 (2007)
5. Neal M.L., Gennari J.H., Arts T., Cook D.L.: Advances in Semantic Representation for

Multiscale Biosimulation. In: Pacific Symposium on Biocomputing, 14, pp. 304–315. WSP, Hawaii (2009)

6. Kirby, M.L.: Cardiac Development. OUP, Oxford (2007)
7. Christoffels, V.M., Smits, G.J., Kispert, A., Moorman, AFM.: Development of the Pacemaker Tissues of the Heart. Circ. Res, 106, pp. 240–254 (2010)
8. Luna-zurita L., Prados, B., Grego-bessa, J. et al.: Integration of a Notch-dependent Mesenchymal Gene Program and Bmp2-driven Cell Invasiveness Regulates Murine Cardiac Valve Formation. J. Clin. Invest., 120, pp. 3493–3507 (2010)
9. Khodiyar, V.K., et al.: The Representation of Heart Development in the Gene Ontology. Developmental Biology, 354, pp. 9-17 (2011)
10. Mungall, C.J. et al.: Cross Product Extensions of the Gene Ontology. Journal of Biomedical Informatics, 44, pp. 80-86 (2011)
11. Balhoff J.P., Dahdul, W.M., Kothari, C.R. et al.: Phenex: Ontological Annotation of Phenotypic Diversity. PLoS ONE, 5, e10500 (2010)
12. Mungall, C.J., Gkoutos, G.V., Smith, C.L. et al.: Integrating Phenotype Ontologies Across Multiple Species. Genome Biology, 11, R2 (2010)
13. Washington, N.L., Haendel, M.A., Mungall, C.J. et al.: Linking Human Diseases to Animal Models Using Ontology-based Phenotype Annotation. PLoS Biology, 7, e1000247 (2009)
14. Smith C.L., Goldsmith C.A.W., Eppig J.T.: The Mammalian Phenotype Ontology as a Tool for Annotating, Analyzing and Comparing Phenotypic Information. Genome Biology, 6, R7 (2005)
15. Gennari, J.H., Neal, M.L., Galdzicki, M., Cook, D.L.: Multiple Ontologies in Action: Composite Annotations for Biosimulation Models. Journal of Biomedical Informatics, 44, pp. 146–154 (2010)
16. Cook, D.L., Mejino, J.L.V., Neal, M.L., Gennari, J.H.: Composite Annotations: Requirements for Mapping Multiscale Data and Models to Biomedical Ontologies. In: 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2791–2794. IEEE, Minnesota (2009)

Appendix: Glossary of Terms

AEPC	Association for European Paediatric Cardiology
AVC	Atrioventricular Canal
BFO	Basic Formal Ontology
CBML	Cell Behavior Markup Language
CBO	Cell Behavior Ontology
CellML	Cell Markup Language
CheBI	Chemical Entities of Biological Interest
CL	Cell Type Ontology
EHDA	Edinburgh Human Developmental Anatomy
EMT	Epithelial to Mesenchymal Transition
EQ	Entity-Quality
FieldML	Field Markup Language
FMA	Foundational Model of Anatomy
GO-BP	Gene Ontology Biological Process
GO-CC	Gene Ontology Cellular Component
GO-MF	Gene Ontology Molecular Function
HPO	Human Phenotype Ontology
MathML	Mathematical Markup Language
MP	Mouse Phenotype
MRI	Magnetic Resonance Imaging
OBO	Open Biomedical Ontologies
OFT	Outflow Tract
OMIM	Online Mendelian Inheritance in Man
OPB	Ontology of Physics for Biology
PATO	Phenotype and Trait Ontology
SBML	Systems Biology Markup Language
VPH	Virtual Physiological Human
XML	eXtensible Markup Language