Multiscale information modelling for heart morphogenesis

2010 J. Phys.: Conf. Ser. 238 012062

(http://iopscience.iop.org/1742-6596/238/1/012062)

Journal of Physics: Conference Series **238** (2010) 012062 doi:10.1088/1742-6596/238/1/012062

Multiscale Information Modelling for Heart Morphogenesis

T Abdulla[1], R Imms[1], J M Schleich[2] and R Summers[1]
[1]Department of Electronic and Electrical Engineering, Loughborough University, Loughborough, UK.
[2]LTSI Signal and Image Processing Laboratory, University of Rennes 1, Rennes, France

E-mail: T.Abdulla@lboro.ac.uk

Abstract. Science is made feasible by the adoption of common systems of units. As research has become more data intensive, especially in the biomedical domain, it requires the adoption of a common system of information models, to make explicit the relationship between one set of data and another, regardless of format. This is being realised through the OBO Foundry to develop a suite of reference ontologies, and NCBO Bioportal to provide services to integrate biomedical resources and functionality to visualise and create mappings between ontology terms. Biomedical experts tend to be focused at one level of spatial scale, be it biochemistry, cell biology, or anatomy. Likewise, the ontologies they use tend to be focused at a particular level of scale. There is increasing interest in a multiscale systems approach, which attempts to integrate between different levels of scale to gain understanding of emergent effects. This is a return to physiological medicine with a computational emphasis, exemplified by the worldwide Physiome initiative, and the European Union funded Network of Excellence in the Virtual Physiological Human. However, little work has been done on how information modelling itself may be tailored to a multiscale systems approach. We demonstrate how this can be done for the complex process of heart morphogenesis, which requires multiscale understanding in both time and spatial domains. Such an effort enables the integration of multiscale metrology.

1. Introduction

After the successes of genomic research in elucidating greater detail on the smallest building blocks of life, the grand challenge remains to make this meaningful to human physiology and medicine. We cannot reasonably expect to uncover the multifaceted ways in which our 23,000 protein-coding genes interact, and affect higher level processes, solely through experiments on animal models. This requires, in addition, a special type of computational modelling paradigm. This should emulate the systems being investigated by providing interaction between processes taking place at different levels of scale, whilst allowing for the reuse of components in different ways.

A number of groups around the world, with the common aim of a multiscale biomedical understanding through physiological modelling and integration, are evolving into a global research network. Multicellular organisms reuse the same cell types and the same signalling mechanisms to produce myriad different functions. Similarly, the Physiome and the Virtual Physiological Human (VPH) research community collaborate to provide curated model repositories of biochemical reaction networks (in Systems Biology Markup Language (SBML) [1]) and biophysical mechanisms (in CellML [2]). These then have the potential to be reused, in whole or part, by other modellers working on different problems, potentially on different platforms, in different parts of the world.

In parallel, the OBO Foundry [3] is developing a suite of reference ontologies that can be used for annotating a wide variety of biomedical knowledge resources. These include images (regions of which can be annotated by segmentation), database entries, publications, computational models and simulation results. OBO ontologies provide increasingly good coverage of biomedical concepts at different levels of spatial and temporal scale (see figure 1). They are becoming increasingly sophisticated in the way they divide the biomedical domain, and at providing a structure that avoids logical inconsistencies. It is not surprising that multiscale modelling efforts have adopted ontologies as a means to integrate with other data resources, which may provide source or validation data. However, at present specifically developed terminologies are used for annotating both SBML and CellML models [2]. The open modelling paradigm is now becoming so influential that some publishers now advise depositing a model in the BioModels database as part of the journal publication process [4].

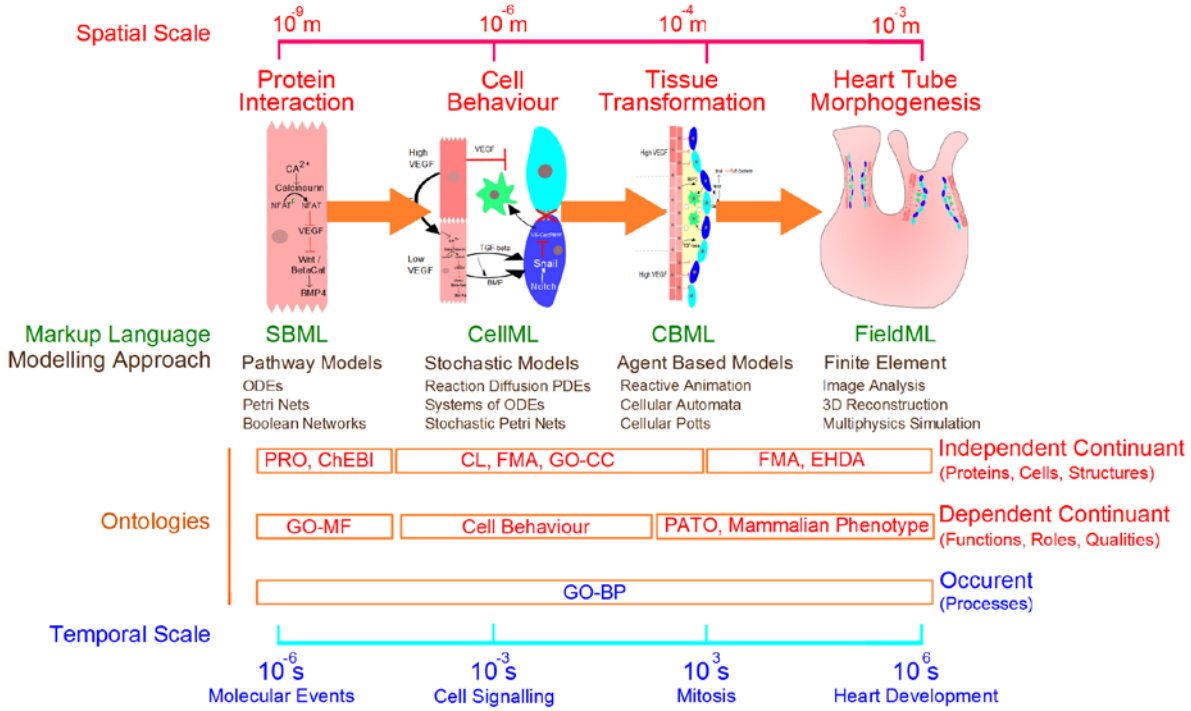

Figure 1. Spatial and temporal scales of the heart morphogenesis multiscale modelling initiative. The modelling framework encompasses spatial scales from 10^{-9}m (proteins) to 10^{-3}m (the primitive heart tube), and temporal scales from 10^{-6}s (molecular events) to 10^{6}s (weeks of development). Modelling approaches suited to each level of scale are indicated, as well as markup languages that aid in the sharing of such models between platforms. All acronyms are defined in the following text.

Physiome modelling efforts have tended to focus on the physiology of adult organ systems, such as the heart [5] or lungs [6]. A multidisciplinary collaboration between clinicians in Rennes and Paris, and biomedical systems engineers and information engineering specialists at Rennes and Loughborough is tailoring the framework for morphogenesis of the embryonic heart. By providing validated multiscale models of morphogenesis processes, the aim is to illuminate the mechanisms by which Congenital Heart Defects (CHD) arise. Many genetic factors have been identified, however what is unknown are the mechanisms by which altered genetic, protein and cellular signalling leads to specific abnormal development. The first cases are focused on the protein interaction and cellular levels of scale, where existing models may be leveraged. Later this will be correlated to phenotypic data of higher level morphology, including images of structures in CHD.

At the bottom of figure 1, we illustrate the reference ontologies applicable to different levels of scale. These are further split between occurents, independent continuants and dependent continuants. These terms are explained in section 3. The multiscale approach is illustrated for epithelial-to-mesenchymal transformation (EMT), a tissue transformation that underlies early development of the heart valves and septa, essential to proper morphogenesis. EMT depends on cell behaviour, which may be perturbed by genetic factors that underlie protein interaction. SBML and CellML are well established, while FieldML is being developed. Cell Behaviour Markup Language (CBML) is still under active discussion, alongside the Cell Behaviour ontology. If successful these languages will fill an important gap in the ontologies required for integrated modelling. This could allow for the same kind of sharing, that is successful for pathway modelling, to be applied to models of cell behaviour.

2. Heart Development

Embryonic heart development commences in week 3 and is completed by week 6 of gestation. The process is quite well documented [e.g. 7]. In normal development, the primitive heart tube first forms in week 3 of development, from the fusion of right and left endocardial tubes. During week 4, it folds back on itself, and the conotruncus rotates approximately 30 degrees short of a complete juxtaposition. While this is happening, a spiral septum grows within the conotruncus, dividing it into the pulmonary artery and the aorta (figure 2). Incorrect rotation of the conotruncus is responsible for a range of CHD, due to improper placement of the aorta and pulmonary artery. Tetralogy of Fallot is a complex disease, defined by four defects: overriding aorta (displaced to the left), pulmonary stenosis, ventricular septal defect and right ventricle hypertrophy. A hypothesis we are currently investigating is that the Tetralogy is in fact a 'monology', with the latter 3 defects occurring as a consequence of the first, caused by insufficient conotruncal rotation. A rarer anomaly, Persistent Truncus Arteriosus, is a complete failure of outflow tract Septation. Septation, rotation and valve formation are all closely interdependent, which to some extent explains the overlap between disease classifications, which in part correspond to slightly different degrees of conotruncal rotation (see figure 4).

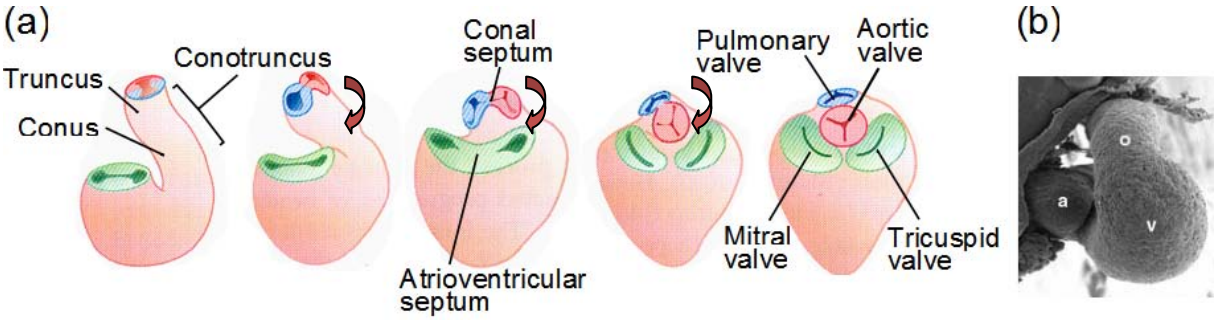

Figure 2. (a) Cardiac looping and aortic wedging (week 4 of development); posterior (rear) view. The conotruncus or outflow tract (OFT) moves such that the conal septum comes into alignment with the Atrioventricular Septum. Simultaneously, septation and valve formation take place via endocardial cushion growth. Normally the OFT rotates about 150° and the aortic valve settles between the AV valves [7] **(b)** Scanning Electron Micrograph (SEM) of a chick heart tube, right lateral view, showing positions of the outflow tract: o, atria: a, and ventricles: v [8]

While heart looping is taking place, endocardial cushions grow in restricted regions of the atrioventricular canal and the outflow tract, via a process of EMT. These go on to form the valves and membranous septa of the heart (see figure 2). Neural crest cells (NCCs) migrate to the myocardial outflow tract and endocardial cushions, and have been demonstrated as crucial to valve development, septation and conotruncal rotation. Many signalling pathways have been identified as important in heart development, with Notch signalling being one of the most prominent. Notch signalling operates via lateral inhibition between adjacent cells. This leads to mosaic patterns in which cells that highly express Notch proteins surround cells that highly express receptors for Notch. This has been shown to have an effect in many processes of heart development, including a role in NCC migration [9].

3. Information Modelling

In biomedical sciences, a wide variety of measurements are taken at different levels of scale. This includes both clinical data and the results of biochemical, biophysical and physiological research. At the molecular level, gene expression analysis produces information on the cellular and tissue locations of gene products. At the cellular level, measurements generate data on variables such as cell volume, metabolism and electrophysiology. There are imaging techniques, applicable to different levels of granularity; such as Magnetic Resonance Imaging and Computed Tomography that can be compared with images from the Scanning Electron Microscope, gene expression maps and gel electrophoresis.

An increasing amount of the biomedical data generated is being made available via web accessible databases. However, this simply leads to a data silo problem, in which finding relevant information is impractical. The coordination and simultaneous querying of such heterogeneous databases is made possible by the adoption of common information models; most successfully with the Gene Ontology (GO). Despite the name, GO is *not* a classification of genes or gene products. Rather, the aim of GO is to provide a standard classification for the *characteristics* of gene products. Consequently it contains information from multiple levels of scale. When annotating gene products, the interest lies in recording their cellular and subcellular locations, their functions as molecules (e.g. they can act as enzymes, transporters or receptors) and the biological processes that they have been associated with (at any level of scale). Thus GO is split into three separate ontologies: Molecular Function (GO-MF), Biological Process (GO-BP) and Cellular Component (GO-CC), as shown in figure 1. The actual classification of proteins is the domain of the Protein Ontology (PRO).

Rather neatly, the three branches of GO fit exactly into the three most fundamental categories of the Basic Formal Ontology (BFO). BFO makes a fundamental distinction between occurents (processes that unfold through time) and continuants (entities that exist in full through a period of time). Continuants are further divided into independent continuants (physical things) and dependent continuants (qualities). These are 'dependent' because do not exist on their own: you cannot have an 'elongated' only an elongated cell or limb. Likewise, there is no 'asynchronous', only asynchronus processes. GO-BP deals with occurents, GO-MF deals with dependent continuants at the molecular level, and GO-CC deals with cellular and subcellular independent continuants (figure 1). It is impractical to define all necessary concepts for biomedicine in one ontology, and for this reason GO is part of a coordinated suite of ontologies: the OBO Foundry [3] with BFO as a basic upper ontology. For example, the Foundational Model of Anatomy (FMA) deals with mammalian anatomical entities, and the Cell Type (CL) deals with the classification of cell types, for all species. GO-BP is one of the only process ontologies, and covers a wide range of temporal scales. This is partly because processes are currently less well defined. This is consequential, in that a focus on morphogenesis modelling involves greater emphasis on the temporal scale, compared to modelling adult physiology.

Two different, but complementary, approaches are emerging in phenotype classification. The Mammalian Phenotype (MP) ontology provides a robust phenotype terminology, which covers spatial scales from cellular to organism phenotypes. This has proved especially useful in providing a manageable means to annotate the rat and mouse genome databases [10]. While MP covers many scales, the degree of granularity is not sufficient for many types of phenotypic data, which are often highly descriptive and specific. The Phenotype and Trait Ontology (PATO) offers a different approach by classifying only fundamental qualities (‘trabecular’, ‘decreased thickness’), without attempting to predefine full phenotypic descriptions. It is then left to curators to compose phenotypic annotations from multiple ontologies, as described further in section 4.

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 modelling projects [11]. This is because the variety of possible classes proliferates as there is a need for more specific annotations. There is clearly a truth to this, and to this extent the PATO postcomposition approach is necessary for a fully integrated multiscale approach. Furthermore, the Ontology of Physics for Biology (OPB) is intended as a means of encoding physical laws and systems

dynamics that can be used in postcomposion annotation of model parameters and physical measurements [11]. For example, the flow rate of blood in the aorta would be:

OPB:fluid flow is_property_of FMA:blood in the aorta.

The same annotation would be used, whether it is pointing to a model parameter, or an actual biophysically determined flow rate. The OPB has been used as a means of integrating models, in different types of code, from different scales, by first converting each of them into lightweight semantic models [12]. This is currently a cumbersome process; however the postcompositional approach with this ontology of physical laws and processes is evidently part of what is needed for multiscale combination of multiscale models and measurement.

Measurements provide parameters for models, as well as a means of comparing models to biological reality under different conditions. Models, on the other hand, are a means of formalising and refining hypotheses for documented biological phenomena and dependencies. Thus there is a clear need to incorporate both under a common framework. Postcomposition of well defined reference ontologies appears to provide the best means of doing this, with PATO providing the basis of phenotypic description, and OPB the basis of physical properties and dependencies.

4. Postcompositon for Multiscale Annotation of CHD Phenotypes

One of the major challenges is to provide information modelling for phenotype data at each level of scale. Model organism databases do this already using the Entity-Quality (EQ) formalism. In general, EQ associates an entity term from a species-specific anatomy ontology with a quality term from PATO [13]. There are reference anatomy ontologies for all the major model organisms, and for humans this is the Foundational Model of Anatomy (FMA). There are also developmental anatomy ontologies, and in this case the relevant ontology is Human Developmental Anatomy (EHDA). It is straightforward to adapt the EQ formalism for developmental phenotypes. The initial steps in doing this are selecting the ontologies relevant to the domain, as well as the types of sources that might be annotated. For the domain of heart development, this is illustrated in figure 3.

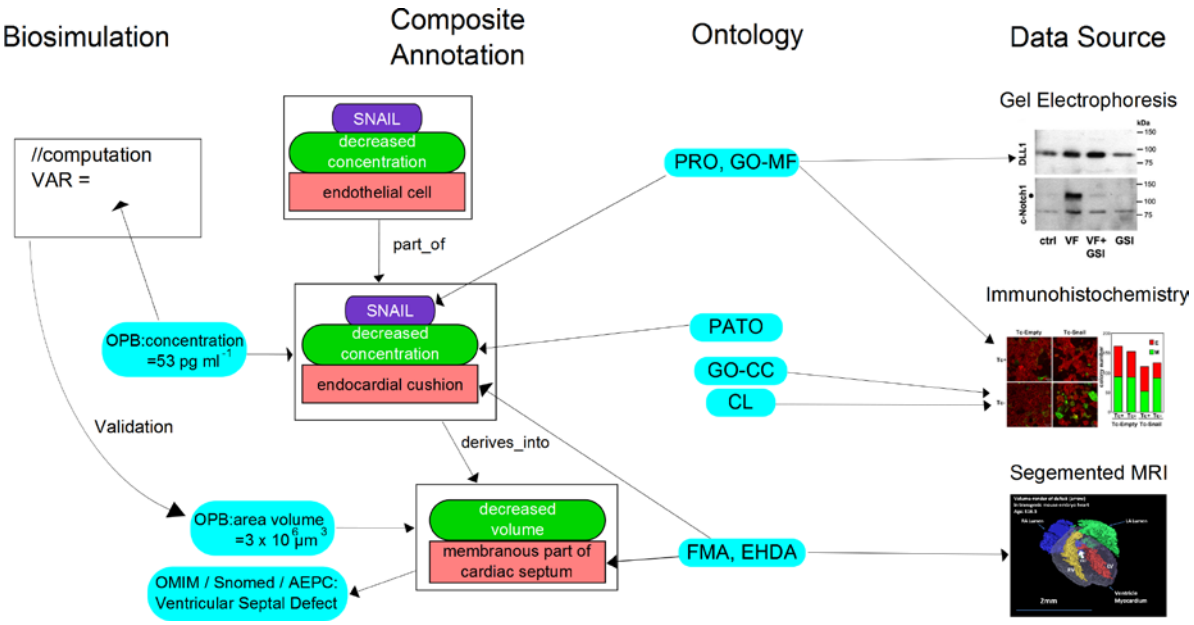

Figure 3. Schema for composite annotation of biomedical data from multiscale sources. PATO allows for composite phenotype annotations allowing for representations such as “endocardial cushion with decreased concentration of SNAIL”, which are composed from multiple reference ontologies. OPB allows formalisation of the physical properties of these composite annotations, such as the concentration of a particular protein in a particular endothelial cell, or the density of mesenchymal cells in an endocardial cushion. These can then be used as variables in a computational model, or as validation of a simulation. PATO composites can be mapped to disease classifications, such as OMIM.

The challenge is similar to the one met in annotating phylogenetic data [14]. There, the importance lies in expressing phenotypes of genetic mutants in the same format as evolutionarily variable phenotypes. The challenge in developmental biology is similarly to link genetic mutant phenotypes to developmentally variable phenotypes. While ontogeny and phylogeny are not strictly isomorphic terms, there are rich and varied inferences to be drawn in fields like evolutionary development. The multiple disciplines to which biological data holds importance further demonstrates the need for flexible mechanisms, which enable reuse and contextualisation. In phylogenetics, one team [15] have greatly increased the workflow of creating EQ representations by customising the open source application Phenote to give it features specifically useful to evolutionary biologists (e.g. a Taxa Panel and a Specimens Panel) as well as entry fields that search specific ontologies as the curator is typing.

There is a need to restrict the largest ontologies so that they focus only on the subtrees relevant to the given domain. The authors have done this by creating a 'slim' of GO-BP, which includes only the processes relevant to heart development. The same approach could be taken to restrict the available cell types (from CL) and anatomical terms (from FMA and EHDA), allowing for the annotation workflow to be further streamlined.

An EQ representation may be defined under a number of categories [15], which we consider here with examples from heart morphogenesis:

- *Monadic states* are those that involve single entities or structures. For example, the range conotruncal CHD are caused by an incorrect roation of the outflow tract. This can be generally annotated as: **EHDA:outflow_tract + PATO:mislocalised_radially**
- *Relational states* are those that describe a phenotype that exists between two entities or structures. For example, an endocardial cushion may be found to have a higher than normal concentration of VEGF-A. This could be annotated as **FMA:endocardial_cushion + PATO:increased_concentration + PRO:VEGF-A**
- *Composite states* involve multiple phenotypes for a single state, which may be monadic or relational. For example, Double Outlet Right Ventricle (DORV) is where both the aorta and the pulmonary artery arise from the right ventricle. In most cases this is also associated with ventricular septal defect (DORV Fallot type). DORV Fallot Type could be annotated as: **FMA:aorta + PATO: associated_with + FMA:right_ventricle**, **FMA:pulmonary_artery + PATO: associated_with + FMA:right_ventricle**, **FMA:ventricular_septum + PATO:dysfunctional**. This is a complex of two relational states and one monadic state
- *Quantitative states* describe a measured value for a variable feature (e.g. size, area, count). For example, bicuspid aortic valve is a condition in which the aortic valve has two leaflets (instead of three). This would be annotated as: **FMA:cusp_of_aortic_valve + PATO:count="2"**

5. Discussion

With postcomposition, there is a lack of exact consistency in annotations between different annotators [13]. This is not always a major problem because, with sufficient guidelines, the differences are usually of specificity (e.g. did they use the FMA term 'endothelium', 'endothelium of endocardium' or 'endothelium of aortic valve'). This is still semantically valid, but the use of coarser terms means a degree of information loss, to be avoided where possible. The availability of the correct tools, such as a customised verison of Phenote, and restriction to terms of a specific domain improves consistency.

However, it is still possible to have different perspectives on the same reality; for example one decision might be whether we are interested in the decreased volume of the membranous septum, or the fact that this means it is dysfunctional. From the perspective of exact quantification, the actual size is important, whereas from the more general disease classification, the interest lies only in the malfunction. To some extent this can be generated mechanically, for example a segmented Magnetic Resonance Imaging can generate volumetric data automatically.

There are often precomposed terms in existing ontologies, which could also be made by postcomposing terms from multiple ontologies. For example, in the Mammalian Phenotype (MP) ontology the term 'abnormal outflow tract development', could be composed as **GO:outflow_tract_moprphogenesis + PATO:abnormal**. The advantages of postcomposition can be demonstrated here, because, in the MP, the only more specific term relating to incorrect rotation is 'transposition of the great arteries' (TGA). This is only one example of a defect which can arise due to incorrect rotation, and in fact there are two different ways in which such a transposition may arise: ~45 degrees in the normal direction of rotation (dextro-TGA) or ~45 degrees in the opposite direction (levo-TGA) (figure 4). The degree of variability possible here demonstrates the advantage of postcomposition: CHD are truly 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: MP classes can be defined as cross products of terms from other ontologies, and this gives the best of both worlds.

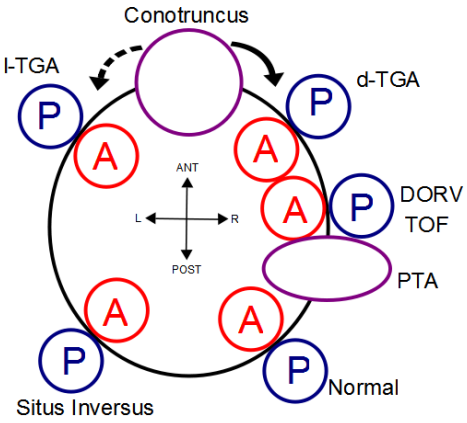

Figure 4. Modifed Van Pragh diagram, after [16], showing the approximate rotation of the conotruncus corresponding to different types of CHD. In PTA, there is no septation into the aorta (A) and pulmonary artery (P). Double Outlet Right Ventricle (DORV) and Tetralogy of Fallot (TOF) correspond to about 90 degrees rotation, while the rotation is approximately 150 degrees in normal development. Situs inversus is a condition where organs develop on the opposite side of the body, and hence the conotruncus rotates counterclockwise rather than clockwise. This also occurs in levo-TGA (l-TGA).

An intriguing possibility is to map anatomical measurements (such as those determined from CHD specimens) to disease classifications. In the example shown in figure 4, ranges of degrees of conotruncal rotation could be used to classify different states (for example 140-160 degrees could be classified as 'normal outflow tract development' and anything outside this range as abnormal). A similar approach could be taken with the position and sizes of the septa, valves and walls of the heart.

6. Conclusion

The challenge expressed in this paper was to correlate initially cellular and protein interaction models of processes such as EMT and neural crest cell migration with data on higher level morphology. Gene to phenotype (disease) associations are increasingly accessible [17]. These tend to use nomenclatures such as the AEPC EPCC shortlist [18], which uses a surgical or anatomical perspective. While this is useful, the gene to phenotype annotations that need to be made, of relevance to CHD, are not always of a 'one gene to one disease' nature. The reality is that several genes are implicated in several mechanisms, which may lead to one of several diseases [19]. The only practical way to enable such proliferous multiscale annotation is through a postcompositional approach, in which entities from several reference ontologies may be combined on the fly.

Model organism databases are well ahead of human biology databases in terms of having well defined semantics and interoperability, primarily through the gene ontology, but increasingly through the postcompositon of PATO and other ontologies through Phenote. This is partly because it is possible to manipulate model organisms, and thus there is a vast abundance of specimen data. However, the increase in non-invasive, high-resolution measurement techniques, and the falling cost of genetic analysis, indicates that it is timely for human biology to catch up. 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.

We have shown how this can begin to be done in the case of human heart development, from a multiscale perspective that coincides with other biomedical multiscale modelling efforts. The data sources of different types, at different scales have been identified, alongside the ontologies suitable for annotation, modelling methods at different levels, and initial guidelines for composite annotation. This demonstrates a method for creating a link between multiscale measurement and multiscale modelling that would assist in closing the loop between physiological and genetic understanding of cardiac development.

References

[1] Novere N L, Courtot M and Laibe C 2007 Adding Semantics in Kinetics Models of Biochemical Pathways *Proc. 2nd International ESCEC Symposium on Experimental Standard Conditions on Enzyme Characterizations* (Rhein: Beilstein-Institut) pp 137-154

[2] Beard D A *et al*. 2009 CellML metadata standards, associated tools and repositories *Phil. Trans. R. Soc. A* **367** 1845-67

[3] Smith B *et al*. 2007 The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration *Nature biotechnology* **25** 1251-5

[4] Li C, Courtot M, Novère N L, and Laibe C 2010 BioModels.net Web Services, a free and integrated toolkit for computational modelling software *Briefings in bioinformatics* **2** 270-7

[5] Noble D 2002 Modeling the heart--from genes to cells to the whole organ *Science* **295** 1678-82

[6] Tawhai M H, Hoffman E A, and Lin C L 2009 The lung physiome : merging imaging-based measures with predictive computational models *WIRES Syst Biol Med* **1** 61-72

[7] Kirby M L 2007 *Cardiac Development* (Oxford: OUP)

[8] Yelbuz T M, Waldo K L, Kumiski D H, Stadt H A, Wolfe R R, Leatherbury L and Kirby M L 2002 Shortened outflow tract leads to altered cardiac looping after neural crest ablation *Circulation* 106 504-10

[9] High F A, Zhang M, Proweller A, Tu L, Parmacek M S, Pear W S, and Epstein J A 2007 An essential role for Notch in neural crest during cardiovascular development and smooth muscle differentiation *J. Clin. Invest.* **117** 353-63

[10] Smith C L, Goldsmith C A W, and Eppig J T 2005 The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information *Genome biology* **6** R7

[11] Cook D L, Mejino J L V, Neal M L and Gennari J H 2009 Composite annotations: requirements for mapping multiscale data and models to biomedical ontologies *Proc. 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Minnesota: IEEE) pp 2791-4

[12] Neal M L, Gennari J H, Arts T and Cook D L 2009 Advances in semantic representation for multiscale biosimulation *Proc. Pacific Symposium on Biocomputing 14* (Hawaii: WSP) pp 304-15

[13] Mungall C J, Gkoutos G V, Smith C L, Haendel M A, Lewis S E, and Ashburner M 2010 Integrating phenotype ontologies across multiple species *Genome biology* **11** R2

[14] Dahdul W M *et al*. 2010 Evolutionary Characters, Phenotypes and Ontologies: Curating Data from the Systematic Biology Literature *PLoS ONE* **5** e10708

[15] Balhoff J P, Dahdul W M, Kothari C R, Lapp H, Lundberg J G, Mabee P, Midford P E, Westerfield M and Vision T J 2010 Phenex: Ontological Annotation of Phenotypic Diversity *PLoS ONE* **5** e10500

[16] Donnelly L F and Higgins C B 1996 MR Imaging of Conotruncal Abnormalities *AJR* **166** 925-8

[17] Barriot R *et al* 2010 Collaboratively charting the gene-to-phenotype network of human congenital heart defects *Genome medicine* **2** 16

[18] Jacobs J P *et al.* 2010 Congenital heart surgery databases around the world: do we need a global database? *Semin Thorac Cardiovasc Surg Pediatr Card Surg Annu.***13** 3-19

[19] Bajolle F, Zaffran S and Bonnet D 2009 Genetics and embryological mechanisms of congenital heart diseases *Archives of cardiovascular diseases* **102** 59-63