VPH-FET Research Roadmap

Advanced Technologies for the Future of the Virtual Physiological Human

VPH-FET

University of Bedfordshire

Istituto Ortopedico Rizzoli

VPH-FET Research Roadmap

A research roadmap to identify future and emerging technologies necessary for the furtherance of the Virtual Physiological Human initiative.

Written by a panel of international experts and coordinated by the VPH-FET consortium.

Advanced Technologies for the Future of the Virtual Physiological Human

Editors: Marco Viceconti, Gordon Clapworthy – September 2011

SEVENTH FRAMEWORK PROGRAMME

European Commission Information Society and Media

Editors: Marco Viceconti, Gordon Clapworthy – September 2011

Executive Summary

We, the community of experts that contributed to this document (see Annex #1 for a complete list) recommend that the European Commission issues Calls for Proposals in the area of FET Proactive that will target topics related to the Virtual Physiological Human.

We make this recommendation because biomedical research and clinical practice are finding it increasingly difficult to sustain the "rate of discovery" that has characterised biomedicine over the last two centuries. This is largely due to the fact that we are reaching the limits of causal reductionism, which has driven this research field so far. *Causal reductionism* implies that "the causes acting on the whole are simply the sum of the effects of the individual causalities of the parts", that is, if we can understand how the parts work, an understanding of how the whole works will naturally follow.

This excessive attention to the so-called *upward causation* (genes cause the cell behaviour, which causes the behaviour of the tissue, which in turn causes the behaviour of the organs, which subsequently defines the behaviour of our whole body) has for a long time overshadowed the strong evidence that, in most living organisms, upward causation coexists with mechanisms of *downward causation*. This term, introduced by Donald T. Campbell in the context of systems theory, can be summarised by the following statement: "the whole is to some degree constrained by the parts (upward causation), but at the same time the parts are to some degree constrained by the whole (downward causation)". Some of the most challenging research questions in biomedicine require that we investigate the biological processes involved in considering the human body as a whole, accounting for all interactions between the microscopic and the macroscopic levels, but this is not feasible with the methods and technologies currently available. We need to develop a new generation of technologies to make this dream of a holistic approach to biomedicine possible. This future technology is called the *Virtual Physiological Human.*

The Virtual Physiological Human (VPH) is "a methodological and technological framework that, once established, will enable collaborative investigation of the human body as a single complex system". First formulated in 2005, and the subject of several Calls for Proposals in FP7, with the earliest projects commencing in 2008, the VPH vision has been explored so far with particular attention to the so called "low hanging fruits" – specific clinical targets for which the existing core technologies could be developed and adapted to provide an integrative approach.

But the full VPH vision is rather more radical in nature, and while for some clinical targets it has now become possible to develop ICT solutions that provide a certain level of integration in the approaches adopted, there is consensus among researchers that the full-scale deployment of this radical vision will require equally radical technological innovation. The current evolutionary approach to VPH technology development, embedded within individual VPH applications, cannot produce this. Instead, we need to take a revolutionary approach to VPH technology development – to step back, consider the broader ongoing technological needs and address these head on.

To this purpose, the VPH-FET support action sustained a consensus process that involved over 150 experts worldwide, the primary output from which is this research roadmap on Future and Emerging Technologies for the Virtual Physiological Human.

In this roadmap, the panel of experts that participated in the VPH-FET consultation and consensus process delineated what are perceived by this research community as the grand challenges for blue-sky technological research associated with the full realisation of the VPH vision: that of a technology enabling a fully personalised and integrative investigation of human physiology in health and disease by means of computer modelling.

While, the research targets described in this document seem quite different from each other, looking at them from a distance we can see a pattern emerging, a big picture that links these topics into a coherent whole.

Everything that is man made craves for simplicity. Occam's razor epitomises how the human mind reacts when faced with an infinitely complex reality – reduce it to simpler terms. The more we understand about life, the more we realise how far biological systems deviate from the ideal of functional perfection. In fact, processes in a biological system frequently involve a number of unnecessary or inefficient steps or imply a number of redundancies and they always embody a certain level of randomness. In other words, all biological processes are inherently highly complex.

For years, biomedical research has tried hard to reduce these processes to simpler terms, achieving important results in the process. However, it is becoming increasingly evident that such an approach can take us only so far in our quest for knowledge: there are biological processes that cannot be fully explained if we reduce them to their parts;

in these, the whole affects the parts while, in turn, the parts affect the whole, and only by *embracing* the complexity can we achieve a full understanding of the systemic nature of these processes.

Ultimately, this is the essence of the VPH vision: develop a new framework of methods and technologies that make it possible for the human mind to embrace the complexity at the root of many important pathophysiological processes. Given our human limitations, to embrace complexity we need to frame it, build it as collective endeavour, and then tame it for our purposes. Frame, build and tame complexity; these are the three characteristic elements that somehow connect all eight topics discussed in this research roadmap in a larger picture:

- **Framing the complexity:** integrative modelling of human pathophysiology. We first need to create a global reference frame, which will allow us to organise data, information and knowledge that we accumulate about any biological system in a coherent and integrated way. Then, we need to represent, properly and generically, the transition from a healthy state to the diseased condition. Last, but not least, we need to find ways of translating what we observe and learn from observing other species into this global body of knowledge of human pathophysiology.
- **Building the complexity:** distributed and collaborative modelling of human pathophysiology. Human pathophysiology is infinitely complex, and can be tackled only as a collective endeavour. By building a web of predictive pathophysiological models that can be reused and composed freely, we can bring structure to this collective endeavour. Similarly, we can use information technology to radically modify the process of knowledge validation, also a collaborative process, by ensuring that studies involving computational models of pathophysiology can also be reproduced and tested as well as physical experiments.
- **Taming the complexity:** interactive fruition of multi-input multi-output integrative models. However, all of the above would be futile if we cannot employ the integrative knowledge that is being progressively and collectively assembled when trying to solve the problems of humanity. That is, we must also develop truly enabling technologies that will be capable of collecting and integrating disparate data and information somehow related to the health status of an individual, and transforming them into knowledge on how this status will evolve in the future. We need technology to explore this maze of complex knowledge to enable us to identify the proper leverage points or the specific actions that can heal or at least reduce the symptoms and the discomfort of our patients.

These are the goals that we believe the VPH research community should set itself over the coming years. Because of the impact that the development of such technologies would have not only on biomedical research and clinical practice, but also on other knowledge domains, we recommend that the European Commission should fund one or more calls for proposals in the area of Future and Emerging Technologies for the Virtual Physiological Human.

Acknowledgement

The VPH-FET project acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the ICT theme of the Seventh Framework Programme for Research of the European Commission.

30 September 2011

Contents

:

Contents

Introduction

Motivations

The Virtual Physiological Human (VPH) initiative had its origins in the summer of 2005 and, since its inception, its primary focus has been the rapid deployment of innovative methods, products and services that incorporate a highly integrated, multiscale approach within the clinical context. While this focus on the final application is probably a significant factor in the success of the VPH initiative, which has attracted over €200m of public research funding, it has come at a cost.

The VPH vision, defined as "a framework of methods and technologies that, once established, will make it possible to investigate the human body as a whole" is somewhat radical in nature, and while for specific clinical targets it has now become possible to develop ICT solutions that provide a certain level of integration in the approaches adopted, there is consensus among researchers that the full-scale deployment of this radical vision will require equally radical technological innovation.

The current evolutionary approach to VPH technology development, embedded within individual VPH applications, cannot produce this. Instead, we need to take a revolutionary approach to VPH technology development – to step back, consider the broader ongoing technological needs and address these head on. This approach, involving "blue-sky" technological research the impact of which may actually go beyond the specific context of VPH, will require a considerable change in the emphases that we, as a research community, place when defining the future VPH research agenda.

We need to move from a collective thinking that revolves around the question: "what can we do with the available core technologies to bring the VPH vision into hospitals within five years?" to a more expansive perspective in which we ask ourselves "what are the fundamental technological challenges that we need to overcome in order to make the VPH vision come true, without any compromise or limitation?" In this process, we should anticipate both the need to bring into the VPH spectrum technologies and expertise not previously adopted in this context, and the expectation that progress made using these technologies to overcome problems inherent in VPH can provide results that ultimately prove fruitful in a variety of fields beyond VPH.

In summary, we need to identify a set of novel or newly emerging technologies that will assist the long-term development of the VPH, along with possible obstacles/bottlenecks that will constrain their adoption. This is the primary motivation of the project *VPH-FET: Future and Emerging Technologies for the Virtual Physiological Human,* which is a Support Action funded within the FET Proactive strand of the Future & Emerging Technologies priority of the FP7 ICT Programme. VPH-FET began on 1 September 2010, and its major deliverable is the VPH-FET Roadmap – the document before you. After 13 months of intense work, the community of experts who participated in the consensus process have proposed a number of Grand Challenges, which are presented in detail in the following pages.

The VPH Community

The VPH community is still quite young. It was initially formed around the activities associated with the FP6 Coordination Action *STEP: A Strategy for the EuroPhysiome* (FP6-ICT-027642), which concluded in March 2007, but the first EC-funded projects formally linked to VPH started only in the summer of 2008. In general, these projects involve biomedical researchers, bioengineers, computer scientists and clinicians. In these projects, the participants have had to address a significant number of challenges associated with broadening both the scope of the work involved in encompassing an integrative context, and the range of multidisciplinary involvement that this requires.

Thus, it is perhaps not surprising that, given the range of immediate problems faced by the initial VPH projects, the VPH community has had little previous contact with the FET Programme and, collectively, had a relatively low prior knowledge of the type of project that the FET Programme is designed to support. As a result, VPH-FET has had not only to identify the future technological directions but also to alert the VPH community to the means by which the technological advances could be achieved via the FET Programme.

Consensus process

As indicated in the discussion above, VPH-FET had several facets:

- **stimulation:** to awaken, in the VPH community, a common view about the long-term technological needs of the VPH, to motivate the members to embrace the technological challenges involved, and to develop an appreciation of how a successful outcome could be achieved
- **insight:** identify the specific expertise, probably in mathematics, computer science, etc., that is necessary to bring the plans into fruition
- **outreach:** attract people with that expertise to VPH and motivate them to become involved in the VPH:
  - as a source of problems requiring fundamental research
  - as an arena within which fundamental research outcomes can be applied and tested.

The first step in developing the consensus process was to try to establish a core set of issues from which to start the discussion and a small, but active, set of participants who would help to establish firm foundations from which the discussion could be driven forward and expanded into broader participation and a wider range of topics.

As mentioned above, it was important to encourage input from disciplines that had previously not had strong interaction with VPH to give the opportunity for cross-fertilisation and the introduction of novel applications of technology to resolve VPH issues that may arise. Consequently, in addition to experts from within the VPH community, a number of experts from other disciplines were contacted with a view to participating in the roadmap development. We received only two refusals, from scientists who already had heavy commitments that prevented their participation, so we ended up with an initial group of 60 researchers, whose interests covered a wide area.

The experts were asked to suggest the types of novel technology that would help to accelerate VPH progress in the coming years. The returns from the experts were classified according to the type of researcher (technologist, modeller, VPH user) and area of activity (data integration, information integration, knowledge integration, automated testing) to which they most closely related. This produced a structure within which topics which were somehow related were clustered, enabling the subsequent discussions to be undertaken in a manageable way.

The experts were invited to a closed meeting in London in December 2010. The main part of the day consisted of 3 sessions in which an "animator" for the each of the three categories of interaction took charge of the debate to identify a list of topics to be taken forward to the Internet discussions that would follow. In fact, the 26 research topics that emerged from the meeting were subsequently reduced to 21, due to some duplication and overlap.

A discussion forum was created on the VPH-FET pages on the Biomed Town web site and its presence was made public by widespread dissemination. Participation in the debates was open to all. The topics were introduced for discussion a few at a time to retain focus according to a schedule covering the period February to June 2011 that was announced in advance. Some topics failed to elicit much interest, while others produced a lively debate. Some of the original experts encouraged colleagues and contacts who had relevant expertise to join the discussions, which proved very helpful.

Ultimately, 9 topics were selected for detailed discussion at the VPH-FET Conference, *Technologies for the Future of the Virtual Physiological Human,* which took place on 27 June 2011 in London. The conference, which had no registration fee, was open to all. Ahead of the conference, animators for 8 of the topics produced documents, which were circulated in advance to all registered delegates; these were discussion papers to aid the debates that would take place at the conference. The final topic, Uncertainty: Modelling and Visualisation, was one of the last to be introduced and it engendered an Internet discussion which suddenly took off just ahead of the conference, so it proved impractical to distribute a discussion paper for it.

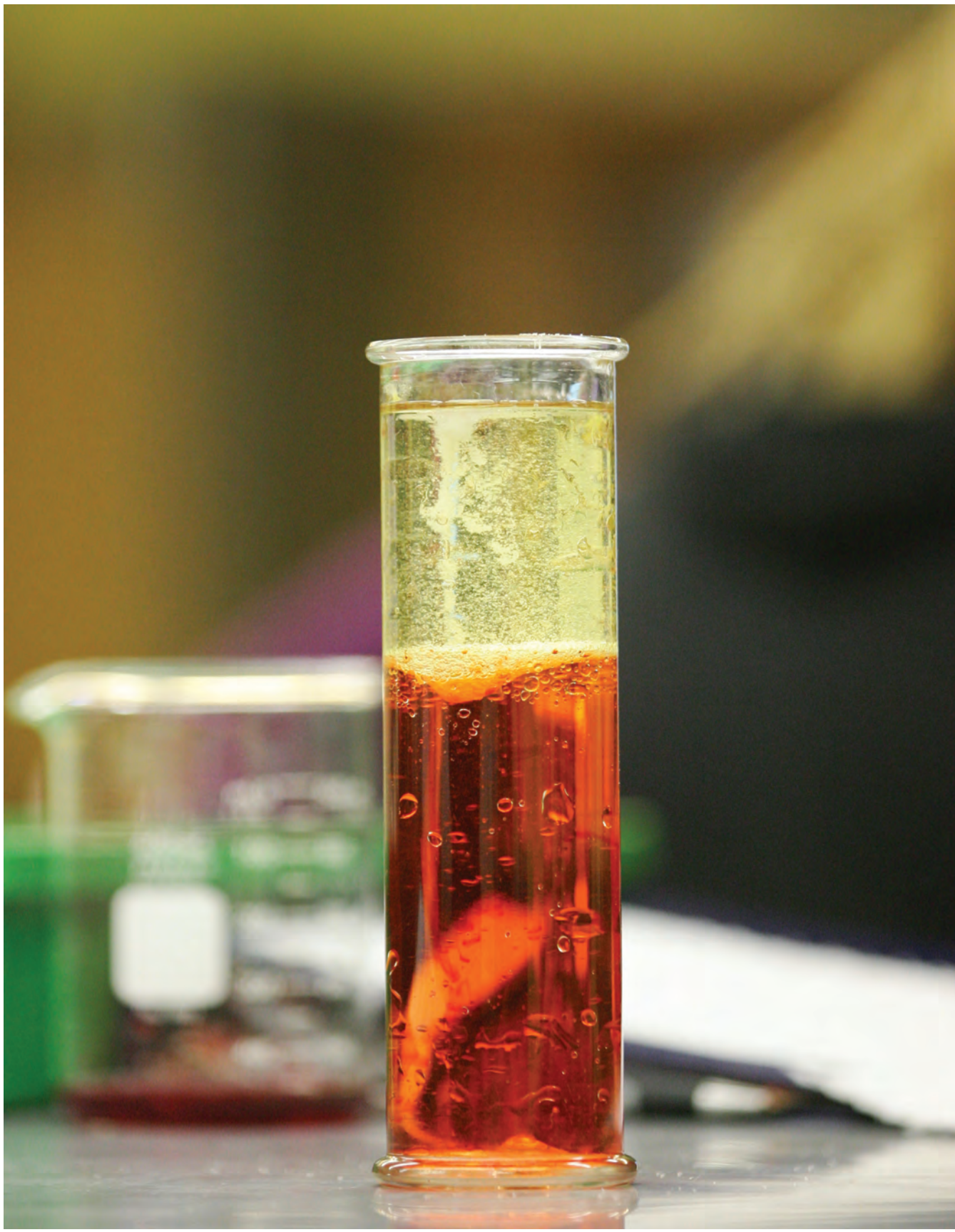
At the conference, each topic had an individual session in which the animator laid the ground in a brief summary and then orchestrated a debate to elicit aspects of the problem that would prove fruitful for future investigation. The session also provided an opportunity to create a small editorial team who would assist the animator in producing an extended article on the topic. These articles form the various sections in the remainder of this document. Unfortunately, it did not prove possible to produce a section on Uncertainty. However, Uncertainty has considerable relevance to visual analytics and to data exploration in imaging datasets, so some discussion of this important topic has been included within Sections 6 and 7.

As the sections were completed, they were added to the draft roadmap and posted on Biomed Town to enable any interested party to make a further contribution if they wished.

Roadmap structure

Sections 1-8 below are the final versions of the articles produced on the various topics that emerged from the VPH-FET discussions as the most significant in terms of VPH-related Future & Emerging Technology.

The final section provides a summary of the features of the previous sections and draws some conclusions about how advanced technology can influence the future of the VPH. A draft Call for Proposals is also included to illustrate how the findings encapsulated in this Roadmap can be brought to fruition within the Future & Emerging Technologies programme of the ICT theme of the European Commission.



Chapter Editors: William Lalinde, Peter Hunter

1

Global space-time reference system for human biomedical data integration

Global

The vision

One piece of pioneering work in this area was the *Visible Human Project*[1]. Using advanced cryogenic serial sectioning methods, a whole cadaver was sliced at 1mm thickness, and high-resolution digital images were taken for each slice. This very large digital data collection was then reconstructed into a 3D stack, which provided the first high-resolution mapping of human anatomy.

The VH project started the field of digital anatomy, which targeted mostly medical education services, such as the Visible Body on-line service[2]. While the project was considered a turning point when the data were released, we rapidly realised that what was missing was very important for most real-world applications. These limitations were summarised in the *Living Human Project manifesto*[3]: "To put it simply – he was only a single human, and he is dead. We do not know how he breathed, walked, swallowed, digested – the VH data totally lacks multiplicity and functionality". Today, there is a widespread consensus[4] that the best way to provide virtual humans with these features involves the development of the so-called *Virtual Physiological Human*[5], intended as a framework of methods and technologies that will make it possible to describe human physiology and pathology in a complete and integrated way.

In parallel, the increased sophistication of computer hardware and software has made it possible for synthetic human performers to be used in animations, games, etc. Today, this is a vital ICT research field on its own right, with various top class labs entirely dedicated to it all over the world. Their research agenda is broad: examples, taken from the Virtual Human Interaction Lab at Stanford University[6] include:

- Avatars and Behavioural Modelling
- The Proteus Effect: how the human tends to represent its avatar
- Haptic Communication in Social Interaction
- Facial Tracking and Emotion Abstraction in Communication
- Homuncular Flexibility: remapping neuromotor control
- Diversity Simulation: try to be someone else for some time.

In these areas, attention is focused increasingly on building synthetic humans that can talk, make facial expressions, act, or interact with us. While the motivation here is different from that of the VPH, as these synthetic humans become more sophisticated, so the points of common interest multiply. In order to improve digital characters, we need to make them move, talk, and behave more like real humans, so this leads to the computer modelling of human physiology and behaviour. On the other hand, in order to make a good computer animation of a car racing in the street, you do not need to model the internal combustion of the engine, so we must recognise that. research on synthetic humans is generally more interested in modelling what you can see from the outside, rather than what happens inside the body.

A third dimension, more recent in conception, is that of using a virtual human visual interface to help users to navigate in the maze of data, information, and knowledge that we are accumulating on the human body and its health. We can enlist the *BodyBrowser*[7] prototype that Google Labs has released, primarily as a demonstration of the capabilities offered by HTML5 and WebGL, of which Google is strong supporter, or *vpHUMAN*[8], a visualisation from anatomical to genetic models in an open portal guided by the VPH framework and evolution of taxonomy standards. Now we can also extend this to BodyMedia and BioDigitalHuman. Other related aspect is the work being done by Antonio Criminisi group at Microsoft Research Cambridge, UK, as part of the *Inner Eye*[9] project, which aims to develop automatic segmenters for 3D medical images, which could be the basis for a search engine dedicated to the human body.

The VPH motivation

The VPH perspective on this matter is focused on the problem of integrating heterogeneous data, information and knowledge. In this light, the aforementioned research lines converge toward a common Grand Challenge that is the definition of a *global reference system for human biomedical data, information, and knowledge integration – namely, the Global Reference Body (GRB)*.

If we have such a global reference system, in principle every single bit of data, information and knowledge on the human body that is shared in digital form can be searched, browsed, interactively explored, for a broad range of reasons, including leisure, education, healthcare, etc. In particular, with respect to biomedical research, the possibility to find, retrieve, and reuse all of the data, information and knowledge that is available on an organ, its functions and its physiological and diseased states, could truly revolutionise the field, increasing exponentially the potential users for VPH technology.

The challenge: Global Reference Body

During the discussion three aspects emerged to compose the challenge:

- **dimensionality:** how many dimensions should this "global space" have?
- **user interface:** how can we provide an interactive visualisation of such a complex space?
- **knowledge management:** how should the data, information, and knowledge be categorised, organised, annotated, etc, to allow it to be usefully and automatically exposed in such a global space?

Dimensionality
This is probably the most challenging aspect. At the current state of our understanding, such a system should consider the following dimensions:

D1) **Space.** We must account for how the body is organised in space and what references are being utilised (model or actual).

D2) **Time.** All physiological and pathological processes need to be described in the context of time.

D3) **Idealisation.** Initially this was proposed as a Scale dimension, but later revised as Idealisation of what is viewed at each scale. Key points are the transformation of what is visualised in each space dimension and the adoption of different idealisations to describe the same process at each scale, accounting for other processes that could not be observed at the next (higher/lower) scale.

D4) **Population.** At the origin of this axis there is an average generic human, then there are all average humans that we obtain by clustering individuals with similar attributes (male, high blood pressure, over 70, etc.), and then each single individual in the world.

D5) **Alterations.** Possible alterations of space based on changes in the body anatomy, accounting for changes due to interventions or body modifications.

D6) **Treatment.** These are the possible actions that we (or our carers) can undertake to transform our health status.

User interface

D1) **Space.** While the amount of information and its complexity is considerable, the anatomical (spatial) representation of the human body and the user interfaces to properly interact with it are already at the centre of an intense research activity worldwide. Current solutions appear effective in a number of cases, in so far as we limit the interaction to a single dimensional scale.

D2) **Time.** Interactive visualisation of time varying quantities is also quite established as a research topic, again in so far as we limit ourselves to a single temporal scale.

D3) **Idealisation.** The user interface issue is considerable, but to a certain extent it maps back to the problem of providing multiscale interactive visualisation, which is being explored by the MSV project[10], at least with respect to space-time dimensions. A further aspect to consider is the role of reference ontologies that represent body structure in informing anatomy-driven user interfaces. Also, O2 Health Link, a Spanish company, has announced the *Activa Central portal*, which is intended to be a unique entry point to databases and tools on biomedical informatics, in this exceeding the current scope of Google Body Browser.

D4) **Population:** How can the difference between the multidimensional model of a set of individuals be defined in a mathematical manner usable in real time? No solution exists at this time that considers a continuous approach, e.g. defining the human face, body or any part of it as a continuous function through time and space. In such a case, morphing from one model to another would not be anything more the calculating their functional differences and creating a continuous transition function between them. However, it should be possible to address this problem also in the discrete domain (time and space not being continuous, but well defined at certain points and moments) supported with lower-dimensional data (such as images of the respected model/object).

In this case, our problem becomes more specific – how to morph from one model to another if we are able to compute the differences between their 2D representations? This solution is evolving and requires agreed 2D perspective references, volumetric data and models, but as always, the challenge is to take into consideration the end user's available hardware resources so that execution to take place in an acceptable level of "real time".

This approach also does not reflect the process of ageing, and with regard to the impact of morphing one part of the model on the model as a whole, further research still has to be conducted, with different "rules" having to be established and defined (partial morphing and interpolation of model parts, e.g. neck when modifying the head). With respect to this, the VPH-FET experts established a contact with Roni Zeiger, developer at the Google Corporation of the Google Body Browser. Roni invited the panel to submit new specifications for future versions of the application. Thus, while it might be sensible to provide specifications for these endeavours, as we develop the rest of the VPH-FET roadmap, it does not seem necessary to include this aspect in the roadmap itself. Still it might remain relevant with regard to the ancillary problem of what users will want VPH data to provide and what type of interface will then be needed to allow them a suitable form of access.

Knowledge management
Research on the use of explicit knowledge representation (such as ontologies) to manage VPH resources is well underway within the VPH community. Collaborations between the VPH NoE[11] and the VPH RICORDO[12] project, for instance, aim to provide the VPH community with tools to effect, share and reason over resource annotations that map to an ontological representation of the body. This work sustains interoperability operations, based on the criterion of anatomical relatedness, ranging from automated dataset

matching to model merging and managing complex simulation workflows.

Challenge description
If we accept the pyramid of knowledge approach we used to partition the discussion topics also here we can separate the problem of building a global reference system in three parts:

Data
Biomedical data involve a huge degree of heterogeneity, in type, accuracy, resolution, information content, etc. Most of the biomedical data generated by imaging or instrumentation, or derivable by processing these data can be represented by a single super-type defined as a posed (positioned and oriented) and bounded multi-scalar field, where location, boundary and field are time varying.

a. **Space.** The general problem is to transform the super-type or any sub-type to and from the space in which it was generated (by measure or computation) to the global reference space. These transformations are called *registrations*. The operational scenario imagined here would require that every time we add a dataset to the global collection, this is transformed to the global space, and the average representations are updated to include also this new dataset. Since manual registration (based on manually annotated anatomical landmarks) is a tedious procedure, there is a vast amount of literature on automated methods. In most work, registration is formulated as an optimisation problem. Numerically, a spatial coordinate transformation is estimated that minimises some "cost function" measuring the residual misalignment of the data. While there is a huge body of literature on the registration of specific data types, currently existing automatic registration methods usually rely on many user-defined parameters, such as the choice of cost function, transformation model and optimisation method, which may have a large impact on the results. Configuration of these parameters is often more art than science and requires a deep understanding of the underlying methodology. A great challenge for the VPH-FET community is the development of a registration method that automatically configures itself based on data characteristics (noise, texture), shape and size of structures of interest, and performance constraints regarding computation time and accuracy. Having such an adaptive registration framework will enable us to circumvent the subjectivity associated with manual parameter optimisation.

b. **Time.** The same operation in time is usually called synchronisation. The problem is conceptually similar to that of registration, but it has specific operational issues, for example on the best way to interpolate over time poses or deformations (changes of the boundary).

c. **Idealisation.** At the first glance, one could argue that data do not contain idealisations. The discussion could become very philosophical, as we notice that the way we observe the data always implies a degree of idealisation; but these reflections would probably take us far from the point. In practice, data generation methods are being developed that have such high resolution that they span two or more scales. In this sense, we can imagine the space being partitioned into consecutive scales, with only a portion of the dataset being exposed at each scale.

d. **Population.** Primary data should always refer to a specific individual. However, a very large body of data available in the literature are provided not as single values, but as average values over a population. In these cases, specific procedures are required to reconcile these "average" data with the single values. As this begins to be applied to personalised cases, the use may become more sophisticated and pose more challenging problems.

e. **Care.** The most relevant aspect is to be able to associate datasets that come from the same subject but over different times, during which a specific care pathway has been followed.

Information
The creation of a global reference system for information is primarily related to the existence of a standardised syntactic and semantic framework within which information is expressed. However, some special cases exist.

a. **Space.** A special case is related to concepts that while expressing a space-time association, do so in a semi-quantitative, qualitative or inherently ambiguous way: concepts such as medial, proximal, apical to be used to define anatomical locations by themselves cannot be immediately translated into a precise space-time region in the global reference system; a manual mapping must be performed between sub-regions of the template and these concepts, and such partitioning, must be transformed back and forth against subject-specific data, to preserve the association.

b. **Time.** Frequently, time information is not present in the data, but is embedded in the associated information (i.e. metadata, or patient folder). Such associations must be preserved, whenever possible.

c. **Idealisation.** The type of idealisation used is information. As such, it should map directly on to this dimension.

d. **Population.** Similarly, the subject ID, or the clustering attributes used to define populations are information. There is an interesting element, however. In a scenario of public databases the biomedical information must be provided in anonymous format, which means that the

only association to an individual is based on an ID number that can be reversed back to the true identity only by the institution authorised to treat these personal data, and then only in the cases considered by the local legislation, and when authorised by the subject him/herself. This typically implies a coding scheme (i.e. DICOM) in which there is a unique code for each institution, and then a unique code for each subject examined in that institution. This means that my data collected in two different hospitals could be presented anonymously as if they belong to two different persons. While in many cases this is not a problem, in others losing the fact that these two observations refer to the same person is an essential problem. This points to a general ICT problem, Universal ID. There is a set of candidate technologies (OAuth, OpenID, Facebook Connect) that might be used to build a user-centric universal ID mechanism, in which the user can authorise such reconnection of information.

e. **Care.** If the care information is standardised into guidelines, or even better modelled as workflows, the problem is trivial. However, clinical practice is far from being so uniform; in reality, the description of care received (from the subject information, i.e. electronic health record) or which can be provided (from literature, industry information, etc.) is far from being formalised and encoded, and it usually spreads over many databases, written in a myriad of formats and encodings, etc. Thus, the problem of transforming this information into a global care dimension poses major challenges.

Knowledge

Statistical associations (or causal relationships) between different sources of data and information represent one kind of knowledge: "Atrophy of the hippocampus is associated with Alzheimer's disease", "The use of drug A causes side-effect B", etc. More generally, knowledge can be seen as a mapping between different domains (of different dimensionality): from blood pressure to the distribution of white matter lesions in the brain, from stenosis grade to stroke risk, from genotype to drug response. Ontologies represent a second kind of knowledge, and provide an explicit and independent identification of biomedical concepts and the relationships between them. Considerable progress has been made in developing reference ontologies for key domains in biology, including gene functions and processes, chemical entities, proteins, anatomy and phenotypes.

a. **Space.** As knowledge provides the link between different sources of data and information, it cannot always be pinned down to a single anatomical region. The global reference model should link a certain piece of knowledge to all the anatomical regions involved.

b. **Time.** Longitudinal studies result in knowledge that relates data/information at different time points.

c. **Idealisation.** A causal relationship (pathway) at one level of idealisation may translate in a statistical association at a higher level. For example, tumour cell death induced by a certain drug can be explained at a molecular level, but may also manifest itself as a change in local magnetic relaxation time, as measured by an MRI scan.

d. **Population.** Knowledge obtained on a population level should be translated back to a patient-specific report/diagnosis/prognosis.

e. **Care.** Practitioners should be able to upload patient-specific data/information and, with a few clicks, generate a report that summarises relevant knowledge that was extracted from the patient data, based on the models present in the system.

Addressing the volumetric GRB challenge: Putting physiological data in a Global Reference Body

In this section we discuss the question of how to link physiological data and models into the Global Reference Body based on ideas being developed in the VPH RICORDO project. A key concept is that of 'embedded' or 'material' coordinates. A region of infarcted tissue in the heart, for example, occupies a unique position within the myocardial wall but moves in space during the beating heart cycle. To locate the infarct position relative to the anatomy of the heart, we introduce coordinates that are attached to the tissue (and so move and deform with the tissue). FieldML, an XML markup language standard for describing spatial (and temporal) fields that is being developed under the VPH project, contains all the concepts discussed here.

RICORDO is exploring the combined use of **anatomical ontologies** and **volumetric image data analysis** to provide **patient-specific** information to **VPH modelling** tools. This approach makes use of **statistical radiological models (SRMs)** – 3D representations of anatomical variation among individuals. These models have the crucial property of being able to map, for any individual under investigation, the equivalent anatomical location in the subject's radiological organ structure (for 2D images, 3D reconstructions, as well as time series). The volumetric annotation of such models with CORDO anatomical IDs allows for the automatic labelling of an individual's anatomical features with standard ontological terms. This step creates the opportunity to:

– determine automatically the patient-specific shape and size of a functionally important anatomical feature, and pass that on to otherwise generic volumetric models or organ physiology

– link up with other VPHDMs annotated using the same ontological terms to integrate regional anatomical

modelling with further information associated with that region (e.g. gene expression)

– establish the basis for the community-based annotation of SRMs, and explore the provision of public SRM repositories based on the concept pioneered by the Protein Data Bank (PDB) for X-ray and MRI molecular structures.

The work is part of the wider issue of interoperability across VPH resources, volumetric as well as non-volumetric (Figure 1.1). Whilst we focus on the issues across different types of volumetric resources, questions of how to link these to the rest of a VPH infrastructure are also being considered. In particular, recommendations for volumetric metadata models and standards need to be articulated with the larger VPH Toolkit development activities. The work is building on three existing representative efforts in the areas of 3D atlas frameworks, statistical volumetric models and computational volumetric models. Respectively, these are the Edinburgh Mouse Atlas Project (EMAP), the GIMIAS system and a computational heart model.

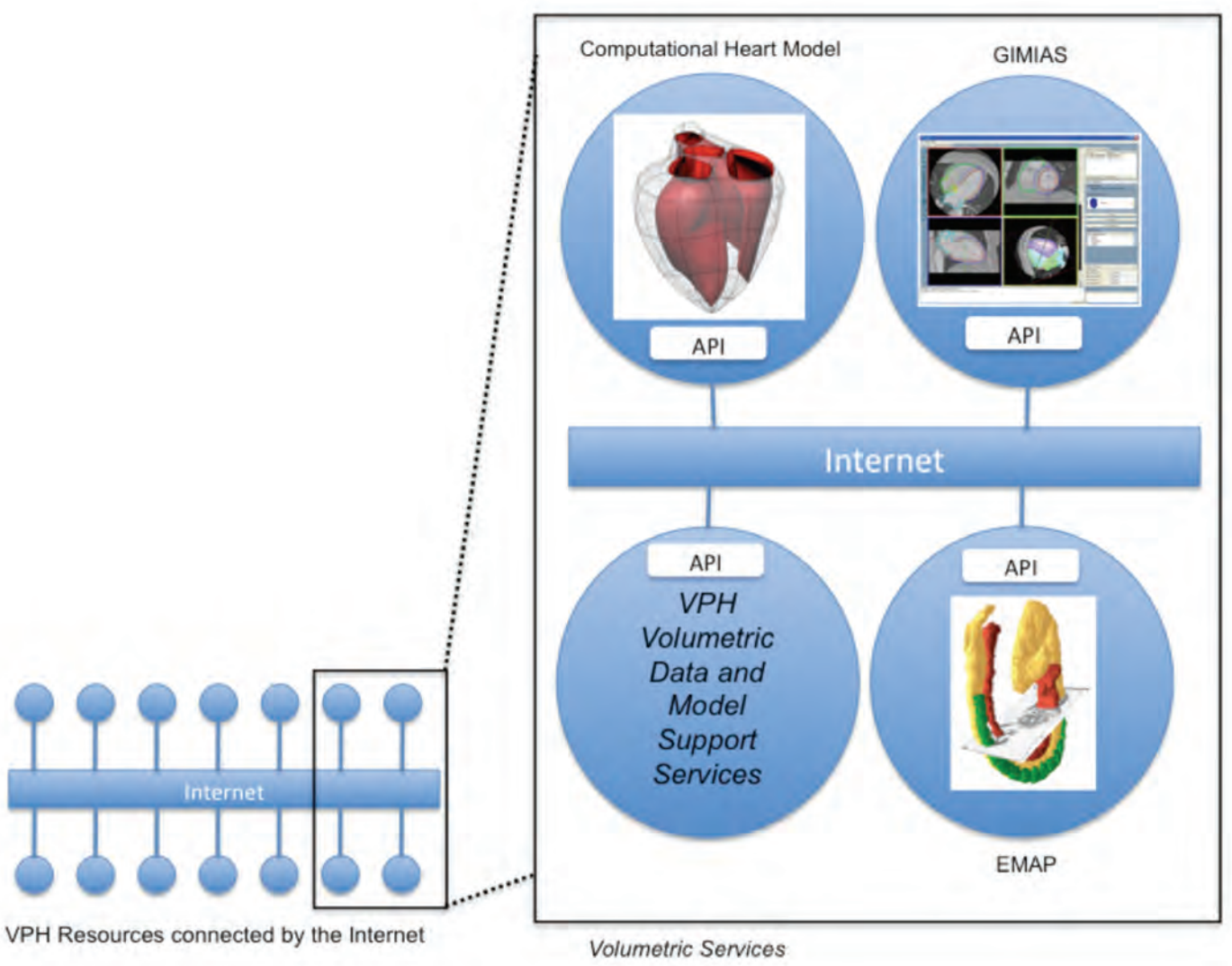


Figure 1.1: Volumetric Data and Models as part of the wider VPH infrastructure (from the Ricordo project application).

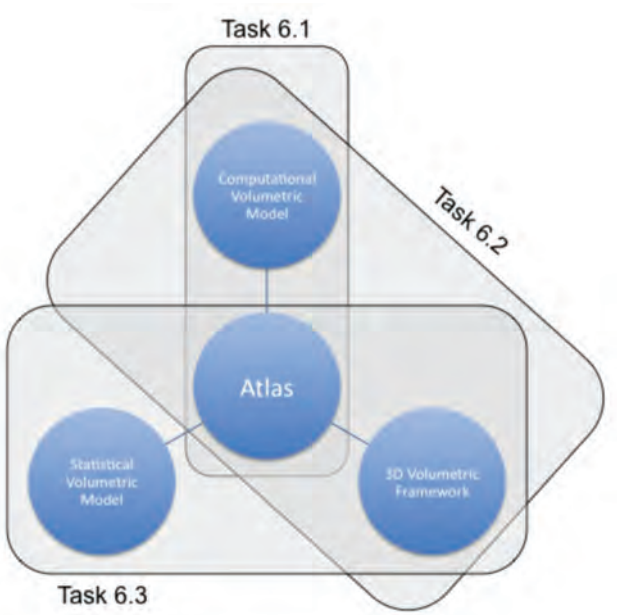


Figure 1.2: Volumetric Data and Models as part of the wider VPH infrastructure (from the Ricordo project application).

Physiological modelling of the heart has resulted in detailed FieldML models associating the computational model parameters within a defined coordinate frame associated with topology and geometry of the heart. This curvilinear model coordinate frame can be mapped on to a standard Cartesian frame (see Task 6.1, in Figure 1.2). From clinical radiological imaging it is possible to build statistical models that represent the normal and abnormal variation of heart geometry within a given population. This model is a series of linked surface models that share node-matched triangulated-surface representations, which enables statistical analysis, for example principal components, to establish normal modes of variation. These then allow, via mapping of a new subject, tests of normality in a statistically significant fashion. In basic biology research, spatial data, such as in situ expression patterns or protein distributions are captured and mapped in the context of models based on a simple volumetric image coordinate framework. These are overlaid with anatomy ontology markup (delineated anatomy) and can be navigated in direct 3D coordinates or via anatomical locations.

These different representations are optimised for specific research and clinical applications; interoperability will enable full transfer of knowledge from the basic research resources through to computational modelling and clinical practice that can produce better understanding of the underlying systems biology, hypothesise targets for treatment and build the infrastructure for translational application of basic biological research. The project is establishing the standards and protocols needed for this level of interoperability by studying use cases selected to establish the mappings required between the different representations. These are modelling the specific behaviour of a patient heart – systems analysis and modelling utilise FieldML models to explore underlying biological processes, and statistical models are mapped to a standard atlas to query anatomy and volumetric-based resources.

The primary use case for exploration of the interoperability issues is based around clinical assessment, physiological modelling and basic biological research into heart function. This is, of course, one of the best developed cases in terms of physiological modelling, and is of major clinical relevance as recent molecular biological studies have linked adult heart function and repair to developmental processes and regulatory networks.

The work is divided into 4 tasks: T6.1-T6.4, the first three of which focus on particular sub-problems, as illustrated in Figure 1.2.

A key question will be whether some form of canonical human atlas or a more abstract conceptual human anatomical space (illustrated as 'Atlas' in Figure 1.2) would be more beneficial for integration of volumetric VPH resources.

We establish embedded (material) coordinate systems in FieldML models that allow the regions of tissue to be assigned FMA[13] labels. This will be done for heart models as an example but it is applicable to all tissues/organs in the body. These coordinates are illustrated in Figure 1.3 for the left and right ventricular myocardium of the heart. The lines shown in Figure 1.3 form three coordinates: h_1 running circumferentially around the heart, h_2 running from the base to the apex, and h_3 running transmurally. Any material point can be labelled by a triplet of embedded coordinates (h_1, h_2, h_3) and any material region, such as the posterior basal segment of the myocardial septum or the right ventricular free wall endocardium, both shown in Figure 1.3, can be labelled by a specified range of these coordinates.

[13] The Foundation Model of Anatomy – an ontology for human anatomy

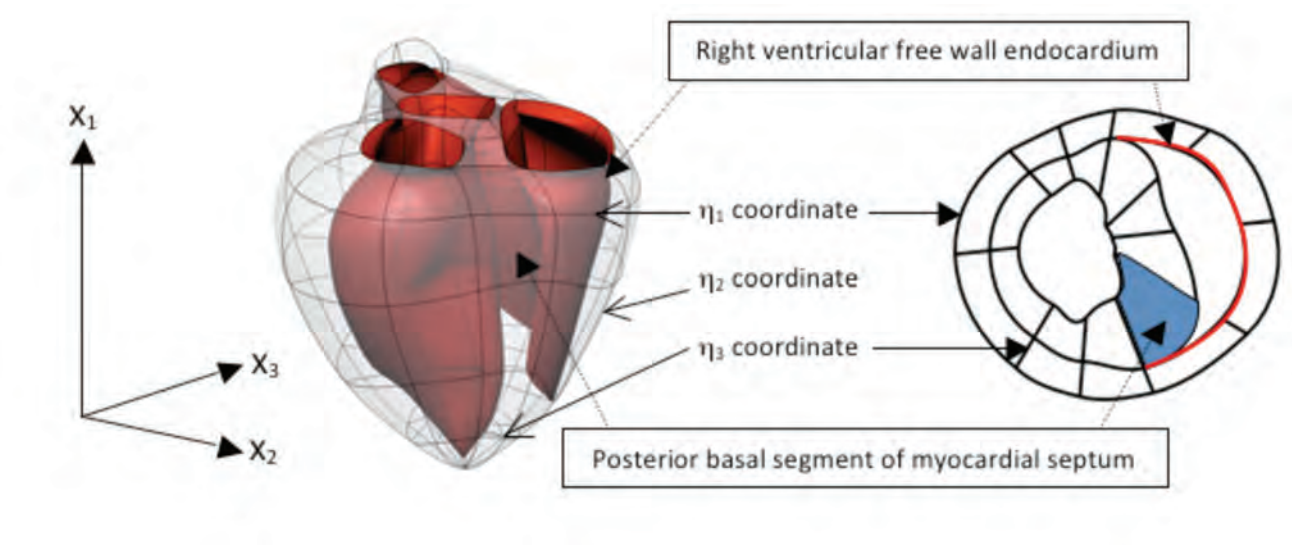


Figure 1.3: Embedded coordinates for the heart. (h_1, h_2, h_3) label material points in the myocardium and move with the heart as it deforms. The red surfaces on the left represent the endocardial surfaces of the right and left ventricles; the diagram on the right is a cross-section of the heart model.

Note that, in order to reference the embedded coordinates for the heart with the embedded coordinates for any other organ in the body, a global Cartesian reference system (x_1, x_2, x_3) is required. This is illustrated in Figure 1.4, where similar embedded coordinate systems are shown for the lungs, skin, skeleton and muscles.

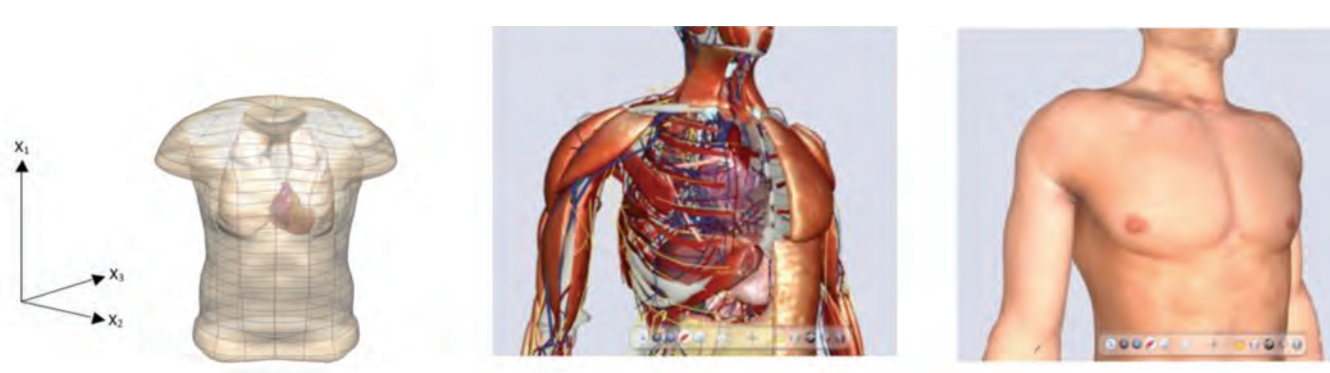


Figure 1.4: . Embedded coordinate systems for the heart, lungs, skin, bones and muscles, shown in relation to a Cartesian reference system in a standardised neutral body position.

Note that once the mapping from (h_1, h_2, h_3)-space to (x_1, x_2, x_3)-space is established for a given organ, other embedded coordinate systems can be established and used for labelling the same points. These mappings must be established in a convenient neutral position for the body, such as that shown on the RHS of Figure 1.4. To assist with annotation, we overlay these coordinates and the associated models with images to which the models are registered, so that the annotations can be applied in relation to image data. Organs are easier to model, while items like nerves and vein vary greatly in position. This is illustrated in Figure 1.5, where the heart model is shown registered to a human MRI image set.

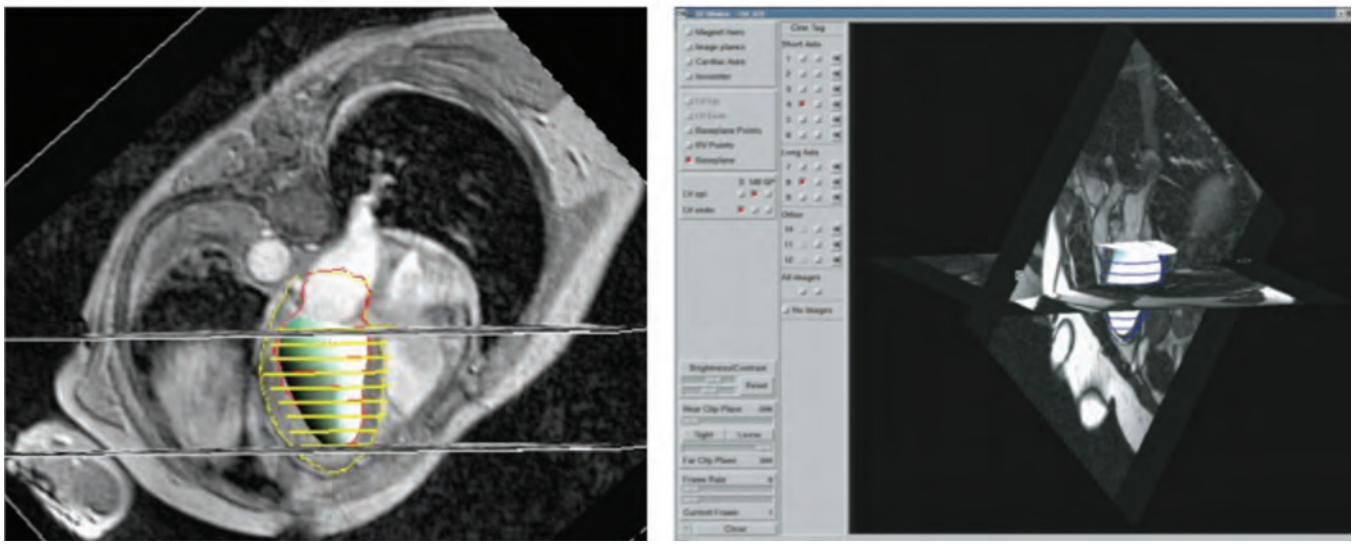


Figure 1.5: Registration of model to image in order to use additional information in the image to help define appropriate labels. A tool for doing this is shown on the right, using the cmgui component of the VPH visualisation toolkit.

Finally, as part of the RICORDO effort, the CellML and FieldML metadata standards are being developed to include model annotation with GO, FMA and other bio-ontologies.

Addressing the GRB knowledge challenge: the RICORDO ontology-based topological frameworks for the anatomical integration of health record data and physiology models[14]

Mechanistic models in the Virtual Physiological Human (VPH) domain describe physiology knowledge in a formal and quantifiable manner. In particular, such models depict physiological processes in terms of the influence that linked physiological variables have on each other (e.g. see Figure 1.6). A significant proportion of variables in physiology models represent either a direct biological measurement, or an inference calculated from these measurements. At the ontological level, most variables correspond to a *quality* (e.g. pressure, concentration, rate of some biological process, etc.) associated with an *anatomical location* (e.g. an organ, or one of its parts).

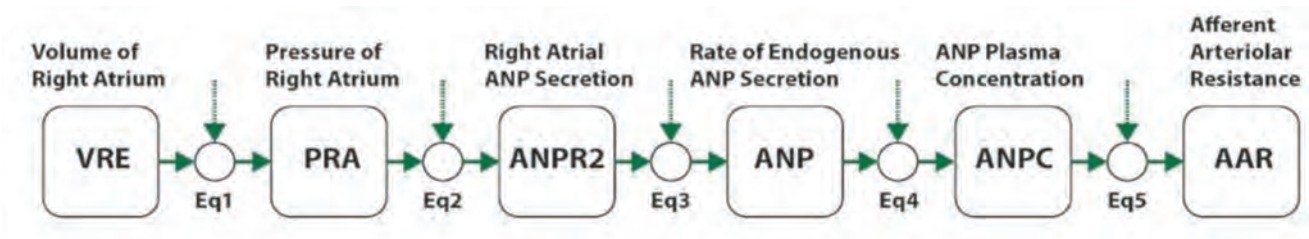


Figure 1.6: An example of a functional dependency series of six variables (rounded boxes) extracted from the Guyton 1992 model, linked via five equations (circles). Independent variables are linked to equation nodes via input arrows, dependent variables via output arrows (therefore, PRA is both an independent variable to Eq2 and a dependent variable to Eq1). Each equation may involve more than one independent variable (hence the dotted vertical arrows). The free-text definition associated with each variable symbol (shown above each symbol) has been copied verbatim from the available documentation from the original model. AAR represents the arteriolar resistance in the kidneys. (ANP: Atrial Natriuretic Peptide).

Equations in mechanistic models formalise the quantitative effect that the value of one variable has on another. In such models, however, provision is rarely made to explicitly depict the anatomical connectivity that conveys the influence variables have on one another. In Figure 1.6, for example, the anatomical route conveying the effect of VRE (a *quality* of the right atrial lumen) on AAR (a *quality* of the kidneys) is not represented per se in the model from which the VRE-to-AAR mathematical relationships were derived. In such models, reference to anatomical location is made indirectly (conventionally via free-text comments associated with physiology variable metadata). The application of an explicit representation of anatomical connections is crucial to a number of key modelling tasks and, in particular, to support:

– the linking of different physiology models to create a single functional network of variables
– the adaptation of a physiology model for a disease scenario – a significant number of diseases may be defined in terms of a set of altered anatomical connections.

Anatomical connectivity is a key organizing principle for a broad spectrum of mechanistic biological knowledge, ranging from long-distance molecular interactions (e.g. endocrine signalling) to developmental biology (e.g. topological operations on anatomical relationships). Establishing an independent ontological framework depicting connectivity relationships between anatomical compartments would therefore support a number of integrative goals in physiology, pharmacology and medicine. In a pharmacological context, for instance, a shared standard for the representation of topological relationships between body compartments would act as a key design template for the construction of pharmacokinetic models, as well as the statistical analysis of anatomically aggregated datasets (e.g. safety and efficacy clinical trial data).

In physiology, a connectivity framework would provide a communal structural basis for the rational organisation of functional (i.e. mathematical) relationships between model variables and an independent compartmental framework on which to integrate individual models. The ongoing effort by the VPH community[15] to annotate datasets and model variables using a communal core set of structural ontologies is therefore crucial in preparing VPH resources to make use of this type of connectivity framework. However, a key second step is required to achieve the above interoperability goals, namely: the consistent application of a multiscale compartmental connectivity framework across the entire core set of ontologies used by the VPH for annotation. The RICORDO project addresses this requirement.

We define compartmental connectivity as follows: two anatomical compartments are connected when a structural (i.e. anatomical) path between them is necessary for the occurrence of a process involving both sites. This core definition of connectivity is therefore established through the ontological integration of biological structure and process, explicitly bridging anatomy and physiology knowledge. Furthermore, the main types of anatomical connections that are essential to convey physiological interactions are classified in this work, based on the above definition applied to the core set of VPH ontologies discussed above. As connectivity categories also take into account biological size scale and topological features, such a framework will provide models with contextual information that is relevant to equations governing the interaction between physiology variables (e.g. boundary conditions, distributions of spatial measurements, etc.).

Furthermore, we describe the use of the above ontology-based framework to model and visualise key anatomical connections and physiological operations. Specifically, we illustrate:

– the application of this approach to integrate location knowledge pertaining to variables in cardiovascular physiology models on to a connected compartmental frame of reference
– the role of this methodology in the development of novel visualisation formalisms in support of the graphical navigation of, and interaction with, VPH models and clinical datasets. To this end, we show the application of this formalism to a clinical scenario drawn from the investigation datasets of endocrine disease involving a number of distinct anatomical locations (see Figure 1.7).

Human anatomy is central to the study and practice of medicine. A significant proportion of clinical symptoms, signs, phenotypes, diagnostic tests and medical procedures are recorded and described with reference to anatomical location. The VPH domain focuses on the mechanistic modelling of a significant number of functional relationships, as well as the statistical correlation of biomedical datasets, that pertain to distinct anatomical locations. Both types of study require an independent and explicit representation of compartmental connectivity that takes into account how physiological effects are conveyed between body sites. The connectivity framework discussed in this work draws upon key reference ontologies of biological structure to provide a widely accessible and scalable method for model integration and data analysis, and lends itself to the visual interaction with biomedical information in a topologically-correct anatomical context.

[14] This section is derived from an abstract presented by Bernard de Bono at the VPH2010 meeting in Brussels.
[15] Through the joint effort of the VPH Network of Excellence (http://www.vph-noe.eu/) and the RICORDO project (http://www.vph-ricordo.eu/

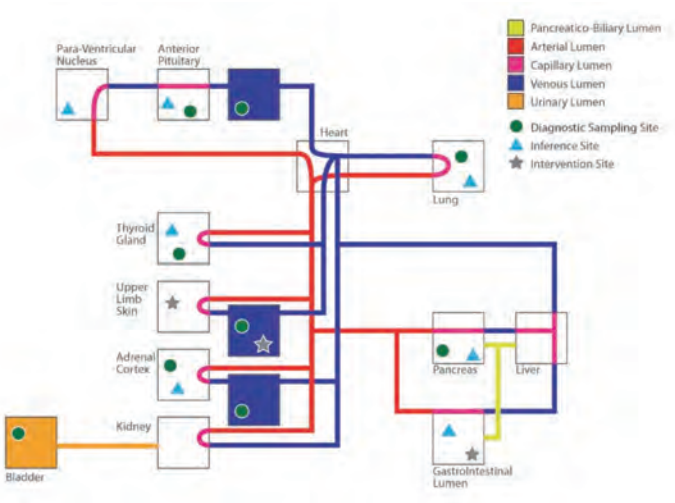


Figure 1.7: A topological graph for the visual organisation of clinical data elements relevant to the investigation of Cushing's Syndrome. In a clinical setting, the wealth of data garnered by diagnostic sampling (green circle symbols) and interventions (grey star symbols) are subsequently collated and integrated by physicians to infer the biological state of key anatomical sites (blue triangle symbols). By mapping VPH models onto the same anatomical connectivity framework that integrates medical data, this work further improves the clinical applicability of VPH modelling.

Impact on Biomedicine

The idea of a global reference system on which we can map all data, information and knowledge about the human body *(Global Reference Body)* could have a huge impact on biomedicine, at least at these three levels:

a. **An interface to data.** Driving consoles in cars, in order to represent all complex information they receive from extensive electronic monitoring available on modern cars represent the information space as… a car. Similarly, we can imagine a generic human body as the primary visual element around which to build the interaction between the users and all the information available. This mostly revolves around advanced interactive multiscale visualisation and human-machine interfaces.

b) **A blender of information.** Information is generally acquired in a fragmented way, and the Global Reference Body provides an integrative representation of such information. Here the accent is on integration, the ability to combine, integrate, fuse information in a synergistic way, and to return such fusion visually to the user, if and when useful. It is about knowledge management, data fusion, image processing, multimodal visualisation and visualisation of uncertainty.

c) **An avatar of our knowledge.** Literally *Avatar* means embodiment or manifestation, and the Virtual Physiological Human is embodied by the Global Reference Body. Here the accent is on the modelling of physiological and pathological processes and their representation in a way that fosters understanding, exploration, and possibly the production of new knowledge.

Then there are three dimensions related to the use case:

a) **Simulation:** for training, retraining, rehearsal, and to support doctor-patient communication we use the Global Reference Body to provide perceptually accurate information about the physiological and pathological processes, as well as of the procedures used to diagnose, treat or monitor the patient. "Perceptually accurate" means that what matters is the accuracy of perceptual (visual, tactile, haptic, auditory, etc.) information exchanged with the user, not the quantification of the underlying pathophysiological processes. The knowledge represented is entirely clinical and completely generic (population-based).

b) **Planning and decision support:** here the Global Reference Body is used to take diagnostic, prognostic or treatment decisions that might involve the precise quantification of some specific indicators (predictors) that are recognised by the clinical user as the most important in driving that decisional process. The predictive accuracy of the models must be good, but it is limited to a few selected indicators. The knowledge represented is mostly clinical, although the precise quantification of the predictors might also require some physical, chemical, and biological knowledge. The knowledge represented is partially personalised, limited to the quantities that directly influence the predictors.

c) **Explanatory medicine:** the Global Reference Body is used as a source of knowledge, as well as a means for generating new knowledge. The Global Reference Body is used to explain why certain symptoms are observed, why a specific pathological scenario will produce certain outcomes, or why a certain treatment will resolve the pathological scenario. The knowledge represented is integral, in the sense that all available knowledge is captured into predictive models, properly integrated, and then exposed via the Global Reference Body. The knowledge is always subject specific, although previous predictions can be used to estimate certain parameters when part of the information is not available.

This 3x3 matrix creates many combinations, when we discriminate per organ system, pathology, or stage of the clinical process (prevention, diagnosis, prognosis, treatment planning, treatment execution, monitoring, rehabilitation).

General Impact

The problems that the creation of a Global Reference Body pose, if solved, could open very interesting applications in many other industrial or scientific sectors where the body of knowledge is large, heterogeneous, and complex.

For example if we imagine the world financial market, where the data are all the financial transactions, the information is in all metadata on companies, funds, banks, financial instruments, etc., and the knowledge is that of economists, but also that fixed by regulatory policies, trade rules, agreements, etc., we end up with a scenario that is not much different, in term of size, heterogeneity and complexity.

Editorial Team

Name	Affiliation
Peter Hunter	University of Auckland
William Lalinde	O2 Health Link
Marco Viceconti	Istituto Ortopedico Rizzoli
Gordon Clapworthy	University of Bedfordshire
Rainer von Ammon	Centrum für Informations-Technologie Transfer
Stefan Klein	Erasmus University
Bernard de Bono	European Bioinformatics Institute
Richard Christie	University of Auckland

Chapter Editors: Ralph Müller, Nic Smith

2

Transforming biological and physiological data across species

Transforming

The Vision

In the introduction to his seminal book "**What is Life?**" first published in 1944, Ewin Schrödinger wrote: "We are only now beginning to acquire reliable material for welding together the sum total of all that is known into a whole; but on the other hand it has become next to impossible for a single mind to command more than a small specialised portion of it."

The acquisition of quantitative and high quality biological data over the last six decades has continued and Schrödinger's insight still remains salient today. Parallel improvements in the performance per unit cost of high performance computing (HPC) and numerical techniques that underpin much of the VPH are now powerful tools for application to Schrödinger's process of ''welding'' or mapping of this biological knowledge.

A key component of this mapping is the translation of insights gained from animal measurements into human relevant contexts. Such translation is not only a fundamental goal of a wide array of life science disciplines but is also, in many cases, an underlying assumption. In traditional wet-lab research this assumption is often ingrained to the point where or how specific animal data can inform human focused clinical practice is not even articulated. While this is clearly an issue, the on-going developments in high throughput measurement techniques and genetically modified animals (in particular mice) are providing new opportunities to address this issue. Specifically, the collection of large and unique data sets both provide the means to, and also increasingly require, this issue to be directly addressed.

More than five decades after Schrödinger's insight, Feri et al. (Mol Cancer Ther June 2006 5:1550) wrote: "Translating modeling results from mouse to human is a challenging problem that is partially addressed by allometric scaling. Some physiologic, species-specific model parameters (e.g., volume flow rates) can be translated across species using the allometric equation, an empirical power law that relates the biological parameter to body weight."

This points to mathematical models being parameterised using animal data, "scaled-up" to the human and then used to predict the biodistribution of the same drug for clinical applications, a theme echoed by many researchers including Geoffrey West and Jim Bassingthwaighte among others e.g. Baxter et al. (Cancer Res. 1995 Oct 15;55(20):4611-22.): "Other interspecies differences that can potentially affect drug pharmacokinetics, such as production and clearance rates of free antigen in blood and binding to target antigen in normal tissues, can be accounted for by making structural changes to the model when scaling from mouse to human."

This citation exemplifies two of three strategies that are currently used to cope with the problem of translating to humans what we observe in animal models:

a) ignore the problem and pretend that a man is a mouse;

b) reduce the problem to an isometric or allometric scaling problem;

c) refactor the model so as to account for different mechanisms observed in the two species, while retaining the same input-output set, which makes it possible still to use the two models as a transformation function.

The VPH motivation

The development of mathematical models using VPH frameworks and tools presents the opportunity to quantitatively assimilate data from multiple sources. Furthermore, because eukaryotic physiology is highly conserved, the structure of existing examples of VPH-style animal-based models is very similar. This means that, using species, temperature and genetically consistent animal models, we may ultimately be able to track parameter changes through the phylogenetic trees to predict human parameters and quantify the relevance of an animal result or disease model to an individual or human population. If this was actually possible, we would be able to tap into the vast pool of data generated for mouse and other animal models to build models of physiology including molecular markers from genomic and proteomic origin. Such information is readily available for all mouse work, including genetically modified animals funded by the National Institutes of Health (NIH) in the United States (http://www.nih.gov/science/models/mouse/sharing/), as NIH-funded investigators are greatly encouraged to make this data public in standardised repositories after initial publication in peer-reviewed journals.

As its name suggests, the VPH primarily targets humans. However, it is clear that there are observations that cannot be made in human subjects, primarily because of ethical limitations. Such ethical limitations are fewer for animals, making these observations possible using a number of animal models. Many features of human biology at the cell and molecular levels are shared across the spectrum of life on earth; our more advanced organism-based characteristics are shared in a more limited fashion with other species. At one extreme are a small number of human characteristics (brain functions and behaviour) that are shared by no other species or, at most, by primates. However, at a lower level, there are characteristics that are shared only with mammals. In this context, the importance of mice in genetic studies was first recognised in the biomedical fields of immunology and cancer research, for which a mammalian model was essential. Although it has been obvious that many other aspects of human biology and development should be amenable to mouse models, until recently, the tools just did not exist to allow for a genetic dissection of these systems.

The movement of mouse genetics to the forefront of modern biomedical research was catalyzed by the recombinant DNA revolution, which began 30 years ago. With the ability to isolate cloned copies of genes and to compare DNA sequences from different organisms came the realisation that mice and humans, as well as all other placental mammals, are even more similar genetically than was previously thought. An astounding finding has been that all human genes have counterparts in the mouse genome which can almost always be recognised by cross-species hybridisation. Thus, the cloning of a human gene leads directly to the cloning of a mouse homolog which can be used for genetic, molecular, and biochemical studies that can then be extrapolated back to an understanding of the function of the human gene. Although the haploid chromosome number associated with different mammalian species varies tremendously, the haploid content of mammalian DNA remains constant at approximately three billion base pairs.

It is not only the size of the genome that has remained constant among mammals; the underlying genomic organisation has also remained the same. Large genomic segments (on average, 10-20 million base pairs) have been conserved virtually intact between mice, humans and other mammals. In fact, the available data suggest that a rough replica of the human genome could be built by simply breaking the mouse genome into 130-170 pieces and pasting them back together again in a new order. Although all mammals are remarkably similar in their overall body plan, there are some differences in the details of both development and metabolism, and occasionally these differences can prevent the extrapolation of mouse data to humans and vice versa. Nevertheless, the mouse has proven itself over and over again as being the model experimental animal par excellence for studies of nearly all aspects of human genetics.

Besides the strong homology in the genome, the mouse is among mammals ideally suited for genetic analysis for several other reasons. First, it is one of the smallest mammals known; second, it has a short generation time – in the order of 10 weeks from being born to giving birth; third, females breed prolifically in the laboratory, with an average of 5-10 pups per litter; fourth, an often forgotten advantage is the fact that fathers do not harm their young and that laboratory-bred strains are relatively docile and easy to handle. Finally, investigators are even able to control the time of pregnancies.

The close correspondence discovered between the genomes of mice and humans would not have been sufficient to drive researchers into mouse genetics without the simultaneous development, during the last decade, of increasingly sophisticated tools to study and manipulate the embryonic genome. Today, genetic material from any source (natural, synthetic or a combination of the two) can be injected directly into the nuclei of fertilised eggs; two or more cleavage-stage embryos can be teased apart into component cells and put back together again in new "chimeric" combinations; nuclei can be switched back and forth among different embryonic cytoplasma; embryonic cells can be placed into tissue culture, where targeted manipulation of individual genes can be accomplished before these cells are returned to the embryo proper. Genetically altered live animals can be obtained subsequent to all of these procedures, and these animals can transmit their altered genetic material to their offspring.

Progress has also been made at the level of molecular analysis within the developing embryo. With the polymerase chain reaction (PCR) protocol, DNA and RNA sequences from single cells can be characterised, and enhanced versions of the somewhat older techniques of *in situ* hybridisation and immuno-staining allow investigators to follow the patterns of individual gene expression through the four dimensions of space and time.

Finally, with the automation and simplification of molecular assays that has occurred over the last several years, it has become possible to determine chromosomal map positions to a very high degree of resolution. Genetic studies of this type are relying increasingly on extremely polymorphic microsatellite loci to produce anchored linkage maps, and large insert cloning vectors, to move from the observation of a phenotype to a map of the loci that cause the phenotype, to clones of the loci themselves.

All of these techniques provide the scientific community with the ability to search for answers to the many questions posed. This will invariably lead to more questions, but the potential is there to elucidate the mechanisms of many diseases and realise effective treatments. The use of one species, such as the mouse, also allows the generation of full physiome maps in a relevant animal model as has recently been performed within VPHOP (www.vphop.eu), an integrated project funded in the 7th Framework Programme of the European Commission. In this project, the effects of anabolic (parathyroid hormone; PTH), antiresorptive (bisphosphonate; BIS) treatment, and mechanical loading on vertebrae in a mouse model for postmenopausal osteoporosis were investigated in longitudinal fashion using in vivo high-resolution imaging. The ultimate goal of this project was to combine all of the results in a physiome map that allows direct comparison of all physiological effects on a standardised time axis. For the purpose of the study, bone loss was induced in 15 week old female C57Bl/6 mice by ovariectomy (OVX). Four weeks of treatment was started either 5 or 11 weeks after OVX, and included a combination of mechanical loading (0N or 8N on the 6th tail vertebra at 3000 cycles and 10 Hz, 3x/week), and either injection of zolendronate (100 µg/kg once), PTH (80µg/kg daily), or corresponding vehicle. An additional group

of sham-operated animals was subjected to the mechanical loading regime. Bone microarchitectural parameters of the 6th tail vertebra were assessed by in vivo micro-CT before OVX, at treatment start, and after 2 and 4 weeks of treatment. To create the physiome map, the values at treatment start (2nd point) were normalised to the basal measurement for each mouse. The effect of the treatment was calculated with regard to the treatment start and the percentage change was scaled according to the main scale of the first measurement. Figure 2.1 shows the map of the trabecular bone volume fraction, but corresponding maps were created also for both the cortical and the trabecular morphometric parameters.

All OVX mice showed a significant bone loss of 24% and 30% at 5 and 11 weeks after surgery, respectively. Mechanical loading resulted in 16-28% of bone gain. While zolendronate stopped bone loss and slightly increased BV/TV, PTH treatment restored the values to the basal level. The combination of pharmaceutical treatment and mechanical loading was additive in the BIS groups while a slightly synergistic effect was found in the PTH early loading group. The physiome map allows a fast overview on the relative efficacy of each intervention and presents a powerful tool to compare the different groups on a single time axis. By including results from future longitudinal animal studies the physiome map can be extended easily.

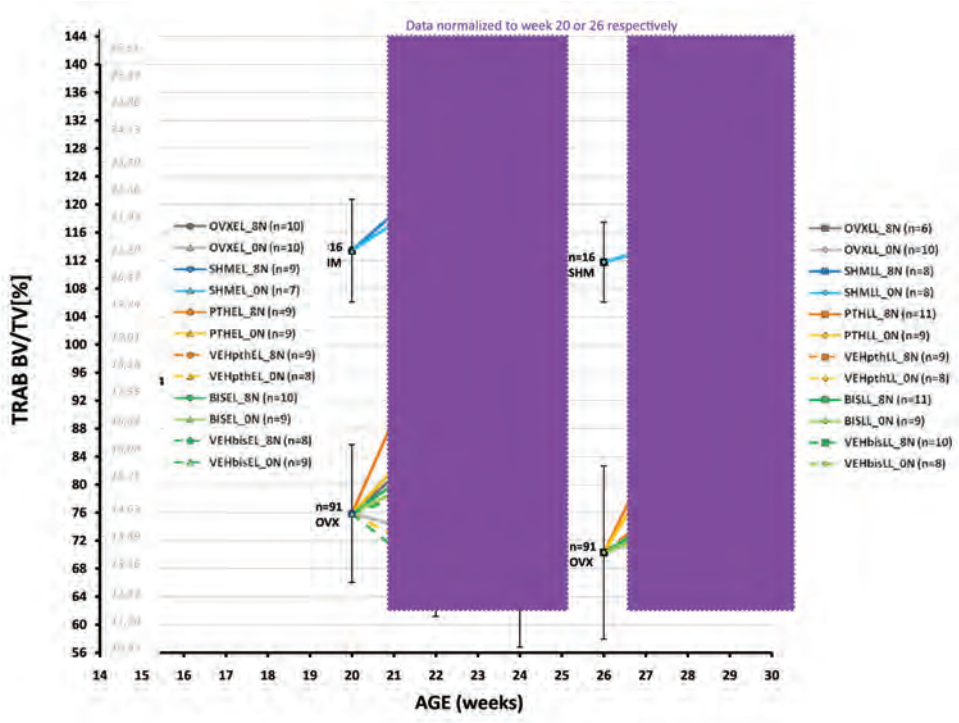


Figure 2.1: Physiome map in a mouse model of human osteoporosis.

The problem is that what we observe in these mouse models needs to be translated into something that will help in understanding human physiology. Sometimes this translation is straightforward, but in most cases it is not; while it may be difficult, it is essential that it is carried out properly. If a model needs inputs that cannot be observed directly in patients, and thus requires the use of estimates from animal models, if we do not perform a correct transformation, there is a significant risk that these data will turn out to be highly inaccurate when used to make predictions for humans.

One option would be to rely more on phylogenetic closeness: so for humans, this would involve the use of primates with a particularly close genetic mark-up as the animal model. Unfortunately, this works only to a certain extent (primarily for the same reasons that motivate the VPH approach to biomedicine). In addition, the use of animal models based on species that are closer to humans tends to raise greater ethical concerns and always involves much higher costs.

An extension of the phylogenetic closeness strategy is to develop multiple models across different genetic strains, populations and species to track parameter variation to both identify missing human parameters and quantify the relevance of animal insights in human contexts. The capacity for mark-up languages such as CellML and FieldML to unambiguously represent and share biophysically based models has the potential to support this process (Terkildsen et al., Experimental Physiology 2008). However, to achieve this goal will require the connection between model parameters and the experimental data on which they are based to be transparent. The maintenance of this link is also important when models are coupled or components reused. This is because this information is key to identifying the sources of model uncertainties introduced, often unavoidably, by combining data from different species or temperatures.

Making model parameter and experimental data links both clear and updatable is thus fundamental to defining the context within which model results can be interpreted in a way that is relevant to the physiological system. This will, in turn, reduce cases in which the failure of a model to represent experimental data reflects an inappropriate structure or parameterisation, thereby increasing the likelihood that such a failure is identifying a gap in understanding, this can, in turn, lead to new insights being gained. For these reasons, a future challenge for the VPH is to develop integrated databases of mathematical and experimental models focused on consistent sets of data in terms of species, temperature and experimental conditions.

Thus, it is important to develop robust and reliable methods to accurately transform observations performed on animals into their equivalent for humans. In much contemporary biomedical research, observations made on mice are directly translated to humans, without any further consideration. This is to forget that animal models are models, no less and no more than computer models. Like any other model they have limits of validity in themselves. Such limits however, could be significantly extended if we could consider not only those observables that essentially do not change from mouse to man, but also those that do transform but according to some transformation law that we can model. In such a case, the model used to translate the observations to the human would become a combination of the animal model and the computer model.

The challenge

Three challenges emerged from the discussion:

a) space, time and space-time transformations
b) parametric adjustment for inter-species integrative models
c) compositional transformations into integrative models

Space, time and space-time transformations

The most evident difference between a mouse and a man is the body shape and posture. In many cases, these anatomical differences are relevant to developing a correct transformation of the observation made on the mouse into the human.

The simplest case is the spatial transformation of a single organ, which in many cases is homothetic. In this case, well-defined elastic registration algorithms are available.

A more complex case is when the transformation involves a change of both shape and posture. This means that we are morphing a complex portion of the anatomy, which is composed of parts that can move relative to each, while remaining connected. For each individual part, the transformation would still be homothetic, but the connectivity negates this condition. Some work is currently being done on the heart and on the skeleton, where global and local elastic and rigid registrations are performed and combined cleverly, but we are far from a general solution to the problem.

A third case is when the transformation, even as a single component, is not homothetic. A typical, though rather trivial, case is when in the skeleton of one species there are two distinct bones, whereas in the other there is a synostosis – a fusion that forms a single bone. Here, no general approaches are available, but some attempts are being made to incorporate evolutionary adaptation logic into the transformation to derive rules for the spatial remapping.

While temporal transformations in themselves are usually not a challenge due to the development of signal processing methods which allow accurate synchronisation, in most cases the combination of space-time transformation (motion) remains quite challenging. In computer animation, there are some interesting methods for what is called "motion re-targeting" which in principle could be explored, and eventually adapted, to solve this problem.

Parametric adjustment for inter-species integrative models

Once a complete space-time transformation is available, we need to transform the observations we make in the animal to the human. If such observations are not merely anatomo-functional or chemo-physical, such a transformation involves a much greater problem – transform the causation.

As outlined above, given the conservation of physiology between species, the causation that links two observables can often be assumed to be the same in both species. In these cases, we can use the same integrative model for both the animal and the human, and the problem is reduced to a multidimensional re-parameterisation.

Central to achieving an effective re-parameterisation is the maintenance of a tight link between model parameters within modelling frameworks and experimental data. This link is fundamental for underpinning and defining the context within which a model component can be used. In contexts in which model components can be appropriately reused without substantial re-parameterisation, coupling methods can be applied to quickly develop new models to study an enlarged system of interest. This powerful feature of component reuse and knowledge capture within biophysically based modelling frameworks is increasingly being applied within the VPH community. However, a difficulty of this approach is that inheritance of model components, due to repeated reuse, now often extends through multiple generations of a given class of models. This has paradoxically resulted in the obfuscation of the link between parameter values and the original experimental data sources. Without clear provenance of sources, identification of inherited parameters requiring updating with more recent measurements is problematic, and the relevance of the model can be eroded.

Thus, as model complexity increases in parallel with available data, processes to identify, assess and critique the basis of model parameterisation must be developed. Specifically, new tools are required to analyse the models themselves to ensure a transparent link between models and experimental data. The application of these tools will be critical in supporting the iterative model development process and guiding a productive cyclic collaboration between experimentalists and modellers. Again, our expectation is that this will reduce cases in which the failure of a model to represent experimental data is due to an inappropriate structure or parameterisation, and this will increase the probability that the failure is a result of a gap in understanding with respect to the specific physiology of the species being investigated. This process will be important for providing a robust modelling foundation for mapping both parameters and results between species.

The incorporation of data from multiple species within VPH models focused on interpreting specific animal experiments or human clinical data remains a challenge in many areas of the VPH. To provide a tangible demonstration of this, Figure 2.2 demonstrates the wide range data sources from a well known model of the cardiac myocyte electrophysiology.

The result is, in many cases, indicative of other models. Once models are fully defined and validated on the animal model, in which both the inputs and the outputs are observable, this is used to predict the outputs (or the inputs, using inverse identification) for the human model, for which the inputs (the outputs) are observable.

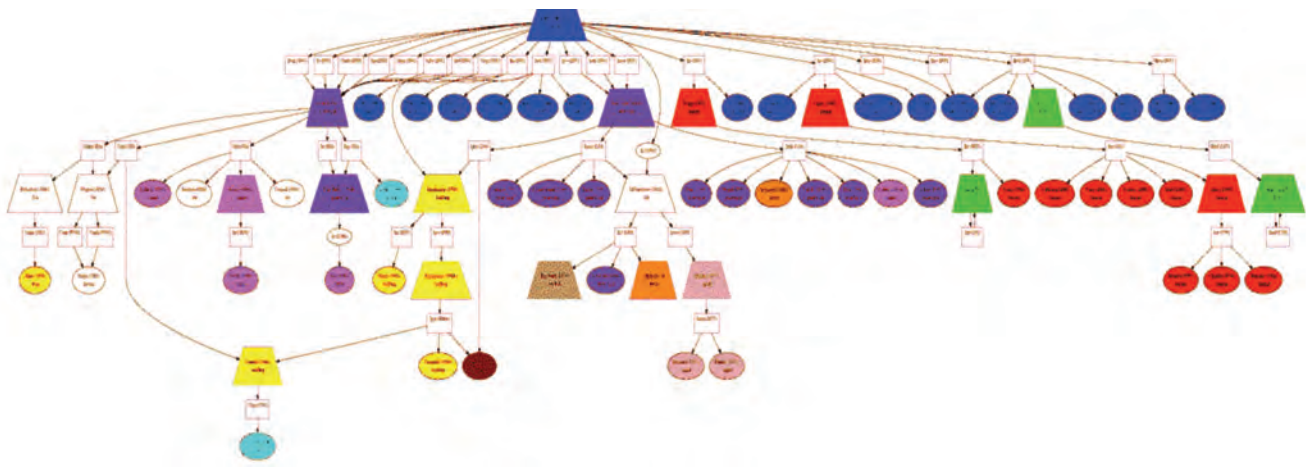


Figure 2.2: The citation (or "genetic") tree of modelling and experimental studies, which were used to define parameters for the Bondarenko mouse ventricular action potential model. Species dependence is defined by colour codes: blue-mouse, purple-guinea pig, pink-dog, yellow-bullfrog, light blue-rabbit, red-human, green-ferret, orange-sheep, grey, light pink-squid, brown-starfish (images from the work of Liren Li).

Compositional transformations into integrative models

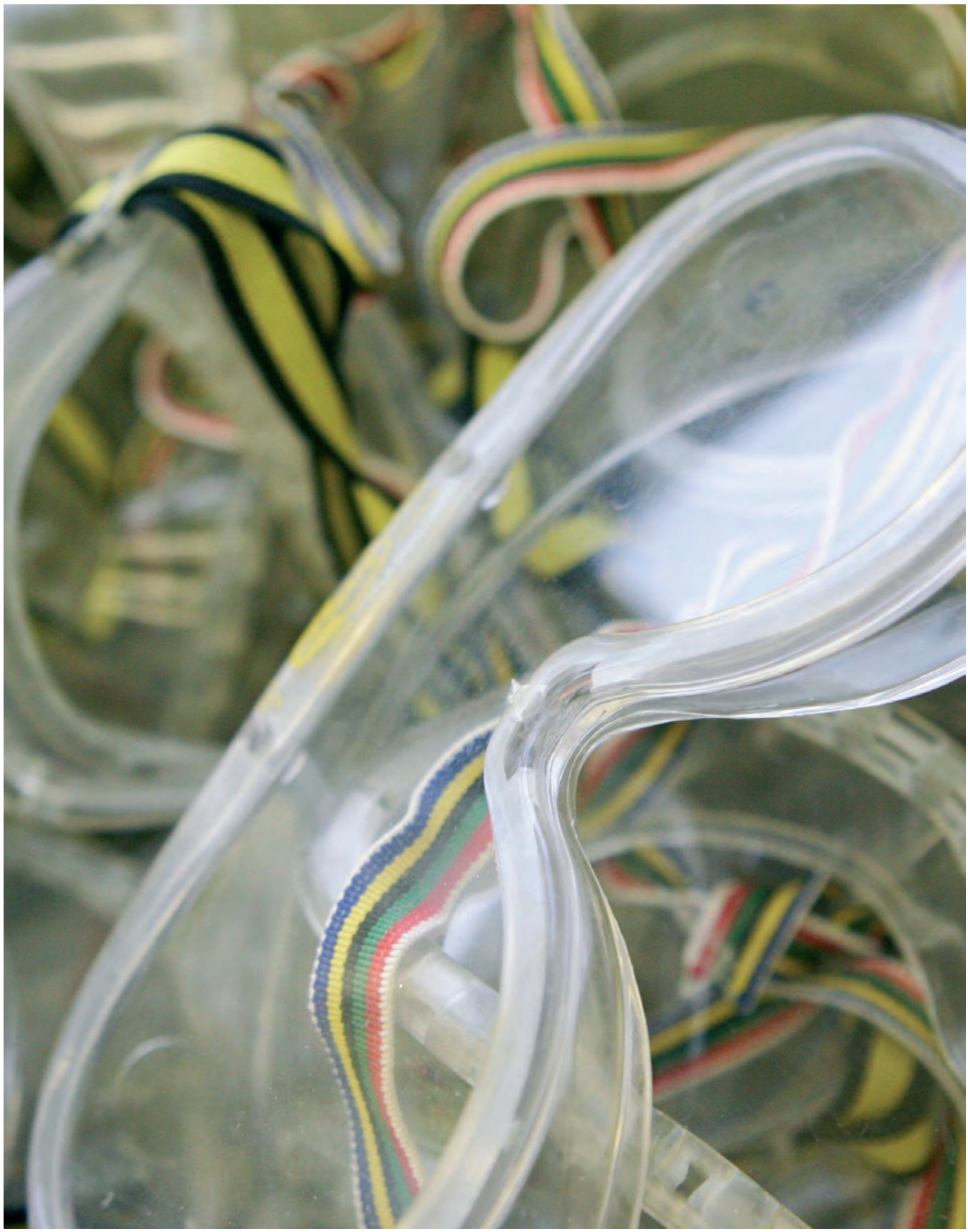
A much more complex problem is when the causation that links sets of observables involves different mechanisms in the two species, i.e. the underlying model is different. This is where we expect the VPH approach to make a greater impact.

Integrative models are predictive models made of component idealisations and of relation idealisations defining the relations between components. In the classic VPH approach, each idealisation is captured in a separate sub-model, so we have component sub-models and relation sub-models, which together form an integrative model.

The idea is to develop an integrative model for each species, using an abductive cycle that starts from the scale at which the phenomenon to be predicted is observed, and adds component sub-models describing processes at larger and smaller scales, until the two integrative models expose the same set of external inputs and outputs, in spite of the fact that internally they are formed by a different sets of component sub-models.

Impact on Biomedicine
The impact on biomedicine would be dramatic. Today, it is quite common to develop, based on animal models, new pharmacological compounds that are expected to produce certain effects, only to find out after long, potentially dangerous, and hugely expensive clinical trials, that this is not the case in humans. Any technology that could reduce this problem, even if only marginally, would have a dramatic positive impact on the pharmacology and medical device industry.

General Impact
The problem of predicting something about a relevant system, based on the observations made on a simpler but similar one, appears of general relevance; it is possible to imagine relevance to applications in fields such as telecommunications, finance, socio-economic modelling, etc



Chapter Editors: Filippo Castiglione, Vanessa Diaz

3

Physio-environmental sensing and live modelling

Sensing

The Vision

The interaction of human physiology with environmental conditions is becoming increasingly important for our understanding of the effect of pollution on our health, but also to link lifestyle with the development of health status in general.

Most sectors of modern medicine, handling either prevention or complex intervention, rely deeply on early, accurate, and complete diagnosis followed by close monitoring of the outcomes, also by means of telemedicine/telehealth technologies. To date, this task is carried out by occasional screening of the individual concerned and producing a (time) series of snapshots at biochemical, mechanical, cellular and molecular levels.

From a biological point of view, although very similar, human individuals show a susceptibility to disease that is unique to them. This simple observation has resulted in the concept of personalised medicines or procedures. However, for a personalised treatment to be really effective, we need accurate individualised information obtained at many levels and in a more continuous fashion, sometimes also requiring interaction with the patient's home environment.

With increasing life expectancy, the number of patients with multimorbidity increases. This complicates diagnosis, prognosis and treatment selection. The traditional medical approach, in which data are collected in a haphazard way, based on inter-current problems regarding a specific diagnosis at the time the patient visit the hospital will not suffice anymore. With increasing life expectancy, the ratio between those who need care and those who provide it, will turn negative. There is an urgent need to shift medical care from institutions to the home environment. More specifically, with the ageing of the European population the number of elders and people affected by neurological conditions such as Mild Cognitive Impairment (MCI) or Alzheimer's Disease is rapidly increasing. The costs of assistance are also growing and they will soon become unsustainable without changing the ways in which these people are supported. To this end, ICT tools are being proposed and studied to improve the support of elderly and disabled people while reducing the overall costs of assistance, but much more is still expected.

The use of technology platforms make it possible to set-up a one-to-many relationship between doctors and patients may be considered as a direct solution for ensuring the necessary quality and intensity of treatment at a sustainable cost. Another required feature of these technological platforms is the quantification of progress related to the subject, which promotes a better modulation of treatment and a faster recovery.

It is clear that this goal has to be accomplished using a sensitive, respectful, non-invasive approach, and that this should avoid excessive interference with the quality of life and, most importantly, should involve the use of affordable and cost-effective solutions.

Much of the world now enjoys unprecedented network speed, high penetration of home broadband and availability of various mobile network options. In this massively interconnected world, in which the social network software industries have paved the way to massive tracking of personal data and the communication technologies have reached the level of consumer electronics and are virtually ubiquitous, it might be possible to wonder how to exploit such approaches to improve medical systems at large.

The questions are: is it possible to develop new hardware-software technologies that are capable of simultaneously sensing physiological and environmental signals, over long periods, minimally or non-invasively, and with a level of comfort that ensures a wide acceptability? Is it possible to process all of these data in real-time with VPH integrative models so as to issue alarms, warnings, or simple recommendations to the subject or to the carers?

In the last decade we have witnessed a rapid surge of interest in sensing and monitoring devices for healthcare and in the use of wearable/wireless devices for a large number of biomedical applications. Also, our environment at home grows more instrumented, interconnected and intelligent [1]. New and more affordable sensor technologies are introducing entirely new monitoring possibilities. Last but not least, in most parts of the developing world, where a large proportion of the global population resides, wireless telecommunication and mobile phones are the definitive means of accessing quality healthcare [9].

Body sensor technology is now becoming available at accessible prices. *Body sensors*[16] are small items of relatively non-invasive equipment that are able to measure biophysical parameters such as the heart rate or the body temperature. Linking these data measurement devices with portable communication systems (i.e., smart-phones) is also technically simple. These advances in technology are enabling smarter, connected personal healthcare systems that can supply crucial information to significantly improve diagnosis, treatment and condition management.

Although the combined use of the above technologies can provide a synergistic effect in activities related to rehabilitation and personal well being, providing at the same time a support for developing new ways of treatment, this approach has so far found little practical application. The main difficulties in simultaneously using these tools are related to their low level of integration; developed by various industrial companies, each device has to be provided with a number of enhancements such as monitoring and visualisation systems and proprietary software platforms that are a physical barrier

to their concurrent use. Furthermore, what is not yet available is a global architecture (or paradigm of data handling at large) for collecting, storing and using this huge amount of data at a level that can potentially be worldwide.

In such a system, the processing of data becomes a potential bottleneck. Nowadays, data are stored in databases and their analysis is mainly done off line, so patients do not directly benefit from information stored in large databases. There is an urgent need for powerful processing algorithms by which it will become possible to integrate and translate large amounts of data into meaningful parameters. These processing algorithms need to be based on models of pathophysiology with continual updates.

The VPH-related vision in this respect is to provide mathematical models (existing or new) that are capable of using this data in a proactive, possibly automatic, manner (hence the word "live" in the title of this chapter). Models that are able to predict the occurrence of a certain event or the emergence of a certain behaviour at an individual or population level (as, for example, in a run-time model checker) would provide an extraordinary instrument for real-time monitoring that would allow reaction and self-adaptation upon an optimised course of action. Analytics programs would monitor device data, collected by sensors and use rules and logic constraints to describe both the environment and the patient health and to compare these against targets, track progress against goals, and send alerts when needed (Figure 3.1). In this way, health-monitoring solutions can become more intuitive, comprehensive and affordable.

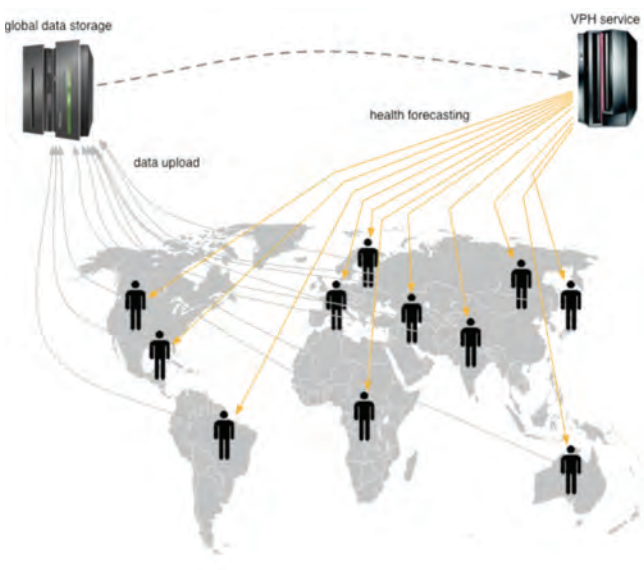


Figure 3.1: Users upload data via mobile network devices. They receive VPH health forecast services through the web or through ad-hoc mobile applications.

We have already mentioned that the European population is ageing, which has the corollary that the number of people affected by neurological conditions, such as Mild Cognitive Impairment or Alzheimer's Disease, and a number of other chronic diseases is rapidly increasing. Examples of areas in which sensor-model integrated systems would prove highly useful are in monitoring:

i) patients with chronic diseases (e.g., Mild Cognitive Impairments, diabetes, epilepsy, chronic cardiac diseases, progressive renal diseases, atherosclerosis);

ii) patients that are hospitalised and need frequent probes;

iii) patients' addiction recovery and long-term drug treatment;

iv) elderly patients in the daily assessment of generic health conditions.

From the perspective of the developing world, personalised applications might not be economically viable, and the prevalent disease pattern also differs. However, the same devices and supportive infrastructure can be modified for both clinical and laboratory diagnosis at the health facility level [10]. An initiative that is working to innovate in this regard is already in existence[17].

As there is a need of flexible ICT tools to support software development in this new application domain, the synergy between VPH models and body sensor technology is straightforward. On the one hand, sensors will feed realistic data-driven models through the collection of data to be used for model parameter estimation and will model result validation in areas in which it has traditionally been difficult; on the other hand, models will be able to assess the impact on the population, optimise the allocation of resources and devise mitigation and containment measures to reduce economic and, more generally, social disruption. Only through a perfect intertwining of the two components will the system will become efficient and effective. The whole vision is based on the idea that with more and better information, people can make smarter choices in managing personal health and wellbeing. This can also aid the design of public health interventions – decision makers will be able to respond quickly to rising disease patterns and intervene smartly, quickly and appropriately.

A crucial feature will be the ease of use and accessibility to data. In fact, the development of this paradigm of data collection, sharing and health forecast, will strongly depend on how easy it will be to share the data and obtain information from the available servers.

The building blocks of such a distributed healthcare system span many areas such as mobile devices, home-based devices, web-based resources, electronic health records and personal health records. Hence, its development will involve

alliances made up from device makers (electronics industry), the healthcare industry, VPH modellers (life science researchers) and ultimately also policy makers to institutionalise its integration into the national healthcare system. An early example of creating a standardised platform for integrating multiple devices for personalised care is that of the Continua Alliance[18].

The challenge

The above vision breaks down into the following key points, each corresponding to a different research, technological, conceptual and societal challenge.

1. Developing a whole spectrum of wearable sensors. What can be measured cheaply and with low invasiveness? Developing home-environment sensors that are cheap and wirelessly connected to personal data hubs. Sensors that can communicate seamlessly with public/private wireless or mobile networks.
2. Developing a data communication system that is secure and allows partial anonymous retrieval to third parties.
3. Developing a robust storage system that is extensible and upgradable. The data collection needs to be organised by using taxonomies that are well accepted within and beyond the VPH community.
4. Developing information systems (such as web servers or internet resources) exploiting such data. This includes VPH methods and models to provide distilled data or predictions from available information.
5. Developing smart and self-adaptive systems, i.e., intelligent environments for monitoring the human health, regulating the uptake of medicaments and predicting individual emergencies. Developing notice network systems based on the overall data that are able to issue warnings to the general population.

Body sensors are already available (see the ICT research stream of sensors and wearable devices), but this remains an area of active and evolving research. Already available are electrochemical, optical and gravimetric sensors, and these allow for measurements ranging from the whole body scale (inertial devices for movement measurements) to the body structure level (textile-based devices for biological signal monitoring), to the so-called bioelectric diagnostic chips that are able to scan bodily fluids for various markers of minor illness and disease[19]. New advanced protein-based sensors are especially interesting as detectors of environmental pollutants (i.e., sensors based on the folding of proteins, peptides and DNA when they come into contact with compounds of interest).

Moreover, implantable *in vivo* monitoring devices are a well-developed area of research addressing problems such as long-term stability and biocompatibility, system integration, sensor miniaturisation, low-power sensor interface circuitry design, wireless telemetric links and signal processing. Apart from technological considerations, a lot of effort in this area of research is devoted to the issue of quality and trust of the service/device. In fact, the level of acceptance of this technology by the users will strongly depend on how reliable, and hence useful, is the final output, and how well the output is used by the research community to improve our quality of life.

The degree of invasive surgery required to implant such devices will finally depend on the type of user. While chronic patients or the elderly are likely to accept anything promising an improvement in quality of life, normal healthy individuals do not. This suggests the most suitable category of citizens to be targeted when seeking to influence public opinion and policy makers at large.

Smart and self-adaptive systems based on two levels of abstraction, logical and physical, can offer in real-time through long-term trend analysis, the prediction, prevention and support of basic daily behavioural and physiological data, building on unobtrusive sensing and advanced reasoning with humans-in-the-loop. The physical level consists of a self-adaptive and self-healing middleware that supports the ensemble of adaptive components and their interactive communication within shared contextual information. The logical level provides tools for automatic reasoning enabling the prediction of spatial/temporal object configurations determining dangerous interactions or physiological damage.

The necessary data communication network is already available, since it can rely on common data/voice network technology plus the Internet, with smart-phones playing the role of the enabling core technology.

In the area of wireless sensor networks, which could provide interesting solutions for the home and pollution detection sensors, it will be necessary to develop wireless protocols and to address the problem of its security as well as problems relating to the performance of large distributed systems, fault tolerance and anomaly detection.

What needs to be developed is a bulletproof communication workflow that goes from the individual to the storage facility in an anonymous and secure way. While, in principle, the data could be stored locally on the device and only later uploaded through a secure connection, in general, embedded systems do not have the capacity to store a large amount of data. Hence the development of secure protocols for uploading run-time measurements is required.

Imagine a physician's tool that could evaluate, in minutes or possibly seconds, a wealth of data from connected health devices plus the complete medical history of a patient and all available medical literature (such as medical records, texts, journals, research documents even ongoing clinical trial results), much of which is unstructured information written

in natural language. This application could suggest possible diagnoses complete with documented "reasoning" or, alternatively, request additional, seemingly unimportant information needed to test hypotheses.

This idea of tracking progress to retain motivation or to monitor chronic conditions and share data with the personal doctor was the original idea behind Google Health[20] which, unfortunately, has been discontinued (end of 2011) because of an unexpectedly low rate of participation in the project. A similar effort (still operational at the time of this writing) is that of Microsoft HealthVault[21].

In data standards, at least two interesting projects are in progress. One is the standard for data storage and communication already developed and adopted by both Google Health and Microsoft HealthVault called the Continuity of Care Record[22] (CCR), the second is Direct Project launched by the US Department of Health and Human Services[23] with the Nationwide Health Information Network initiative in March 2010[24]. The NHS Interoperability Toolkit[25] in the UK and HITCH[26] in the EU are similar ongoing initiatives.

The driving philosophy behind these two efforts is in line with the aim of this topic. In particular, the communication of health information among healthcare organisations, providers and patients is traditionally achieved by sending paper by mail or fax. The development of a standard for data exchange seeks to benefit patients and providers by improving the transport of health information, making it faster, more secure and less expensive. It will facilitate "direct" communication patterns with an eye on achieving unprecedented levels of interoperability.

From the VPH point of view, the development of a general storage system consisting of large data-warehouse facilities in charge of providing controlled access to users, does not itself represent a challenge. However, collected data needs to be organised in a strict but extensible and upgradable manner. This ultimately comes down to the problem of adopting a standard for names and symbols of biological objects and the use of controlled vocabularies and ontologies to describe repository content. This is indeed a VPH outcome to be opted for and capable of fostering further development.

Finally on this issue, the aspects connected to the possibility of combining data and models in a close synergistic effort to create new information in a way that is both accessible, at one extreme, and secure from malicious usage, at the other, is both stimulating and challenging. There are a number of important aspects that should also be considered and safely addressed from an ethical point of view. For instance, data from which epidemiological information at the level of geographical regions can be derived has an enormous strategic value for industrial sectors such as pharmaceuticals. Data security or integrity is most essential especially if cloud computing is being considered.

In summary, data needs to be kept private and secure; it should be shareable with health professionals and downloadable for use elsewhere (and accessible through mobile devices). Data should be organised according to standardised ontologies and stored in digital formats that are well defined and already adopted.

In relation to the foundation and development of mathematical and computational methods for predicting the spread of disease, the system will prompt the development of new (or the adaptation of old) large-scale, data-driven mathematical and computational models endowed with a high level of realism. VPH models enabled by ubiquitous sensor data will allow the forecast of critical events. Moreover, the design and implementation of original data-collection schemes will themselves be motivated by identified modelling needs. Think, for example, of the collection of real-time disease incidence data using innovative mobile sensor ICT applications. The set up of

[17] MoDiSe: http://www.modise.org
[18] http://www.continuaalliance.org/index.html
[19] http://www.techradar.com/news/computing/cutting-edge-intel-s-bioelectric-wonder-chip-425788
[20] Google Health: http://www.google.com/intl/en-US/health/about/index.html
[21] Microsoft HealthVault: http://www.microsoft.com/en-us/healthvault/
[22] A standard proposed by the ASTM (ASTM International, formerly known as the American Society for Testing and Materials). http://www.ccrstandard.com/learnabouttheccrstandard. The Continuity of Care Record (CCR) is a core data set of the most relevant administrative, demographic, and clinical information facts about a patient's healthcare, covering one or more healthcare encounters. It provides a means for one healthcare practitioner, system, or setting to aggregate all of the pertinent data about a patient and forward it to another practitioner, system, or setting to support the continuity of care. The primary use case for the CCR is to provide a snapshot in time containing the pertinent clinical, demographic, and administrative data for a specific patient. (Source: http://www.astm.org/Standards/E2369.htm).
[23] http://healthit.hhs.gov
[24] http://healthit.hhs.gov/portal/server.pt/community/healthit_hhs_gov__nationwide_health_information_network/1142
[25] http://www.connectingforhealth.nhs.uk/systemsandservices/interop
[26] http://www.hitch-project.eu/about

computational platforms for disease forecast and data sharing will generate important synergies amongst different research communities and countries.

A critical problem is how to drive consumers who lack extreme motivation (such as suffering from a chronic condition) to share their personal data. In fact, as has already been discussed, the system should rely on the participation of the population to collect real-time information on the distribution of biological parameters or diseases by means of their personal body sensors and smart-phone devices (Figure 3.2). This is itself a big challenge. How can we reward the individuals who spend their time and money (since cell data connection is not available for free) and who share personal information with distant entities such as research institutions? Is there a real need to transform this system into an economic model in which to produce revenue, or will the promise of having access to a better health system be sufficient?

In principle, the potential savings that live modelling and continual monitoring may provide, through early diagnosis and pre-emptive treatment, may open the possibility of applying novel forms of project financing for innovation. In the same way that many public works programmes across Europe have been financed through a mixture of public and private funding in conjunction with the agreement that the private investors are entitled to a return on their investment through tolls or the equivalent for a sufficient period of time, cost reduction or controlling eHealth innovation may also attract private investment, if a share of the potential reduction in the cost of treating patients can be passed back to the original private investors in the form an "Innovation Dividend". In a contemporary setting, the value of the saving, of which the original private investors would be entitled to a share, could be derived from the reduction in the average cost (adjusted for inflation) of the care of a sufficiently large number of patients with a specific disease within a region that had been selected to trial the innovation in question for a pre-defined duration. This would result in an economic incentive for innovation that could attract a wide variety of healthcare providers, IT companies and investments institutions, whilst initially stabilising (and later on reducing) the costs of healthcare delivery, management and innovation.

In the future, a product of a fully functioning VPH-based innovative patient- and process-oriented care, based on live sensor-derived model-guided medicine and on consequent model-guided clinical workflows, spanning the entire health continuum from prevention to diagnosis and treatment to rehabilitation and nursing care, will provide scientifically justified health reference costs. As a result of the expected increase in early diagnoses and pre-emptive care, the outcomes of such a system could favourably reflect, in terms of cost, on the contemporary average costs system that was described above, enabling both private investors to benefit from a significant return on their innovation investment and for healthcare providers and patients to benefit from lower costs and higher quality of care. In the long run, there is even the possibility that the traditional relationship between income (national or individual) and healthcare expenditure that results in healthcare seeming to be the equivalent of luxury goods could be broken and replaced by a relationship that sees the core costs of healthcare delivery and quality detached from income levels and more closely aligned with innovative solutions to fundamental healthcare needs. Finally, scientifically justified reference costs and evaluated outcomes-oriented management could replace the black-box (hidden, pragmatic) approaches to healthcare systems (including diagnostic, therapeutic, systemic and managerial) and fully exert a role as potential drivers of change.

Another critical issue may be the lag time between data-collection and individual benefit. If meaningful parameters are directly and readily available, this will improve motivation for patients and carers. Meaningful outcomes could be used for online feedback and coaching. Data may be supervised by health workers and summarised and used as guidance during consultation at health institutions. A constraint of minimal lag time between data import and producing output may put extra demands on data-collection and processing. This perfectly underlines the necessity to develop powerful data processing algorithms, based on pathophysiological models that are capable of extracting information at far greater speed than is performed nowadays on static databases.

An inspiring example could be found by looking at the recent societal and economical phenomenon of social networks. These software systems are actually collecting an enormous amount of data without providing any financial reward to individuals. They collect data because people are willing to share information with other people. Note that this is indeed one of the possible reasons for the failure of the GoogleHealth project, as enthusiasm for sharing personal health data possibly requires a relationship with an institutional partner rather than a software company. This further suggests that the involvement of institutions in such a vision is essential for the active support of a critical mass of citizens. The HealthSpace, a personal health record platform operated by the NHS, which is also suffering from the same disappointingly low utilisation, provides a suitable anecdote for reflection[27], suggesting that, direct incentives to the patients, citizens or population are required that go beyond simple institutional support.

It is important to show the data providers (i.e., the patients) what the potential benefits are. On one hand, the system could rely on a kind of "social contract" whereby motivated individuals have a clear return in term of health assistance. On the other, a business model could be adopted to gain from potential market opportunities. The question is whether a system such as the one envisioned will prove to live up to

user/patient expectations or if the whole solution requires a concrete real market opportunity to exploit. Perhaps the answer lies between these two extremes in that sensor vendors and the communication technology industry can exploit a market opportunity, while data exploitation and health forecasting, although curiosity and research driven, will provide enough critical services to boost the interest of a part of society that believes in technological advances within the health system.

With respect to data collection, two interconnected points are at the core of the challenge: how to collect the necessary data and how to ensure that there is no abuse of the data. Both questions need to be handled in unison and robust solutions provided if we want to employ this technology. Also, this scenario will be markedly influenced by the growing use of electronic patient records that will spread in the next few years to cover all clinical activities.

Finally, besides the technical challenges facing body-sensor technology (design, biocompatibility, invasiveness, reliability, energy consumption etc.), there are a number of legal, societal and ethical challenges that need to be addressed.

The VPH motivation

The expertise accumulated in addressing VPH-type (hence system level) modelling problems by VPH projects has great value. VPH will add to the body sensor research agenda the strong connection between models and data. Data collection must be extended with physiological modelling techniques, which are capable of processing the acquired data in real-time and providing individualised warnings and alerts instead of just relying on some simplistic, generic rules.

The task of sensing, collecting and managing the data gains a larger significance when combined with the possibility of producing new and valuable information (i.e., a prediction) on the health status of single individuals or entire populations (Figure 3.2).

Moreover, previous VPH projects have addressed the problem of creating models at different levels of description, which are by definition, quite sophisticated and complicated. In so doing, VPH physiological models rely on a large number of parameters describing a living organism (or a substantial subsystem), most of which we cannot observe in vivo for an individual (at least not without being excessively invasive). Without such data, we shall not fulfil the promise of truly personalised medicine, and even the best models and simulations may become meaningless.

This is where the data collection we have been discussing enters into play, as it addresses the problem of parameter estimation by supplying useful data. The future VPH web service should be able to estimate a number of meaningful parameters and provide valid ranges, taking into account

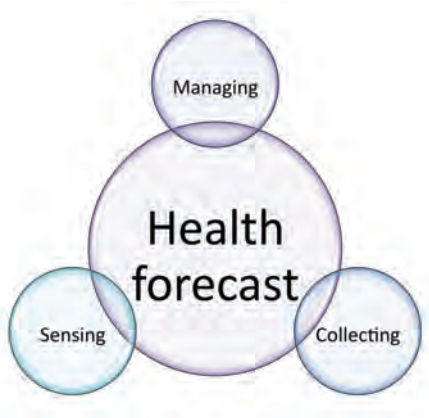


Figure 3.2: The task of sensing, collecting and managing the data gains a larger significance when combined with the possibility of producing new and valuable information on the health status of single individuals or entire populations.

factors such as idiosyncratic characteristics of the patient, its geographical context, regional economic conditions (hence the wealth of the healthcare system), etc.

Therefore, a key challenge of this vision is the combination of environmental and wearable lifestyle sensing technology over substantial time periods, with more healthcare oriented clinical approaches allowing a snapshot of very detailed physiological parameters also of the internal structures and dynamic processes within the body.

Impact on Biomedicine

Revolutions in biotechnology and information technology have produced enormous amounts of data and are accelerating the extension of our knowledge of biological systems. These advances are changing the way biomedical research, development and applications are carried out. Clinical data complement biological data, enabling detailed descriptions of various healthy and diseased states, progression and responses to therapies. It is the availability of data representing various biological states, processes and their time dependencies that enables the study of biological systems at various levels of organisation, from molecule to organism, and even at population levels.

Multiple sources of data support a rapidly growing body of biomedical knowledge, but our ability to analyse and interpret these data lags far behind data generation and storage capacity. Mathematical and computational models are increasingly used to help interpret biomedical data produced by high-throughput genomics and proteomics projects. Advanced applications of computer models that enable the simulation of biological processes are used to

generate hypotheses and plan experiments. Appropriately interfaced with biomedical databases, computational models are necessary for rapid access to, and sharing of, knowledge through data mining and knowledge discovery.

Computational biomedicine will provide the possibility of developing not just qualitative but truly quantitative analytical tools (that is, models), on the basis of the data available through the system just described. Information not available today (large cohort studies nowadays include thousands of individuals, whereas here we are talking about millions of records) will be available for free.

Large cohort data will be available for online consultation and download. VPH integrative and multiscale models will benefit from the availability of this large amount of data in that parameter estimation will be possible in a statistically meaningful manner. At the same time, distribution maps of important parameters will be generated and continuously updated. Through a certain mechanism, the user will be given the opportunity to express his interest in this or that model in order to set up a consensus model selection process. Moreover, models should be open for consultation and annotation.

Positive outcomes of such technology include flexible and user friendly services for the simulation of case studies, testing and validation of specific assumptions on the nature of a given disease, understanding the world-wide distribution of a parameter or disease pattern, the ability to hypothesise intervention strategies in cases such as the spread of an infectious disease, advanced risk modelling, etc.

Applications on the market at the time of writing include a diabetes management solution that combines and analyses data from the patient's insulin pump, a continuous glucose-monitoring device and a blood-glucose meter and makes the results available to the individual's doctor. Having a real-time view of blood sugar and the ability to deliver insulin precisely when needed greatly assist diabetics in reducing the risks associated with erratic sugar levels.

This is a typical example of the use of data which, after processing, provide direct feedback of a patient's health status producing the possibility that the patient can promptly intervene autonomously without any delay or interference from health personnel. The shift of medical care from institutions to the home environment will lessen the burden on the labour force and will help to keep health care budgets within reasonable limits. One of the challenges will be to deal with the increasing complexity of states of multi-morbidity and older age in which it is not feasible to guide interventions based a single parameter. Failure and function of various organ systems have to be taken into account simultaneously, i.e., guidance of blood sugar levels with respect to organ damage (low is better) or cognitive function (higher is better). These kind of choices can only be made based on a thorough knowledge of different clinical phenotypes, data on the state of different organ systems and adequate data-processing algorithms based on pathophysiology models in which disease interactions are taken into account.

General Impact

As we learn from [1], a number of interesting and related surveys have been conducted. We discover that the rapid adoption of mobile interactive devices has provided a viable gateway for consumers to transmit health data. Currently, 17% of mobile phone owners (29% of those with ages between 18 and 29) use their phones to look up health or medical information, while 9% of mobile phone owners (15% in ages 18-29) have smart phone applications that help them track or manage their health [2]. In fact, 10% of all apps downloaded from the Apple iTunes store (1.09 billion downloads) are related to healthcare, medical and lifestyle [3]. One example is the Pfizer Mon Krono Santé application, which serves as a memory aid and offers a personal health record for chronic disease sufferers [4]. Gaming devices are viable conduits, too. Bayer DIDGET, for example, is a plug-in for the Nintendo DS gaming system targeting children with diabetes [5].

The growing number and increasing maturity of web-based resources are providing more opportunities for consumer self-service and peer support. Bayer, for instance, offers a comprehensive support program called BETAPLUS for multiple sclerosis patients [6]. In addition to an application for the Apple iPhone device that assists with injection timing and site reminders, Bayer's solution includes a robust website with educational tools, peer support and access to solution-trained nurses. Also, a product called LUCAS has developed an innovative and low-cost microscopic appendage to a smartphone which is currently being trialled for laboratory diagnosis in Africa [7]. Similar initiatives using microchips or sensors are currently under development by the EU [8].

Clearly, the building blocks are there and gaining traction; but greater value will be gained by bringing the components together to provide a step-change in diagnosis and treatment in terms of both patient outcomes and healthcare system efficiency.

This technology will potentially provide a huge shareable collection of biomedical information worldwide. Devices will be most successful when they provide data that would not otherwise be available because of the measurement frequency required or the need to capture the data in real-time at a particular point in a given physiological process.

This information will be "live", meaning that it will be updated continuously and instantaneously. Data crawlers and

analysers will extract and produce data from data, thus producing interesting distilled information. VPH models will use this data to make predictions. Medical institutions can use this information to set up healthcare services such as monitoring systems, warning systems or aid systems.

Mobile and home-based devices monitor vital signs and activities in real time and communicate with personal health record services, PCs and smart-phones, caregivers and healthcare professionals

Smarter health systems continually analyse information from multiple devices and other sources to derive insights and recommendations for the individual's health regimes.

Two EC-funded projects are already tackling some of these problems at different levels (or scales). At the epidemiological/population level, the FET project EPIWORK[28] proposes a multidisciplinary research effort aimed at developing a framework of tools and knowledge for the design of epidemic forecast infrastructures to be used by epidemiologists and public health scientists (also called "Internet-based surveillance"). This pins down the application of the ideas described above, aside from clinical usage, as of potential interest for epidemiological modelling in times of pandemics or epidemics in developing countries, or during natural or environmental disasters. At the subject/individual level, INTERSTRESS[29] is developing a set of personal system tools and services for the collection, classification and aggregated representation of individual stress patterns.

Along these lines, it might be worthwhile mentioning another prominent example of Internet-based surveillance, which is the Japan earthquake and tsunami warning system that automatically issues alerts via television and cell phones shortly after the first, less harmful, shock is detected, providing time for people to prepare for the more powerful shock that will follow.

It is noteworthy that the VPH-FET vision of the underlying future need focused on monitoring the health status of European citizenship, shares a common view with the "The Future of the Internet"[30], that is, to exploit the true unprecedented connecting power of the Internet thanks to mobile and wireless networking and services in order to provide a bridge between market-driven research and fundamental research to meet Europe's future needs.

On other aspects more related to the development of communication protocols for wearable or implantable devices, it is interesting to look at the coordination action CA-RoboCom that designed and described the FET Flagship initiative "Robot Companions for Citizens" (RCC). This envisions an ecology of sentient machines that will help and assist humans in the broadest possible sense to support and sustain our welfare.

Lastly, the VPH-FET vision can have a significant impact on the objectives of the JADE[31] project. As already mentioned, European citizens are getting older and are increasingly living with chronic diseases because, although their health condition is better than that of earlier generations, they live longer thanks to advanced medical care and therefore end up with chronic conditions and minor disabilities that are often manageable by home care. This has highlighted shared concerns by regional governments about implications for future provision of welfare and health services. This demographic change poses significant challenges to European society and its economy.

The JADE project will develop and promote a Common Research Agenda and Joint Action Plan addressing one of the most promising cluster applications of Ambient Intelligence technologies in everyday life categorised according to the needs of a healthy ageing population: independent living services and Telecare. They embrace eHealth as an enabler for a range of activities: tele-consultations, transfer of records, telehomecare, telehealth and vital sign monitoring, interpersonal communication, remote care and social support (home monitoring, navigation and tracking, etc).

As a final note, the whole vision is truly interdisciplinary and potentially ground breaking. These questions tackle a range of technological challenges; hence its adoption by the European Community will possibly create new skills and ultimately new jobs, one of the VPH community's contributions to Europe 2020.

[28] http://www.epiwork.eu

[29] http://www.interstress.eu

[30] http://ec.europa.eu/information_society/activities/foi/research/index_en.htm

[31] FP7-Capacities n.266422, 2011-2013 – JADE: joining innovative Approaches for the integration and Development of transnational knowledge of clusters policies related to independent of Elderly

References

1. IBM Global Business Services, Executive report, The Future of connected health devices, March 2011, http://public.dhe.ibm.com/common/ssi/ecm/en/gbe03398usen/GBE03398USEN.PDF
2. S. Fox, "Mobile Health 2010," Pew Internet & American Life Project, October 19, 2010, http://pewinternet.org/Reports/2010/Mobile-Health-2010.aspx
3. M. Lyall, "Health apps contribute 10% to 10 billion app downloads," Healthy Living, January 22, 2011, http://www.knowabouthealth.com/health-apps-pitch-in-10-to-speedy- 10-billion-downloads/7664/
4. D. Tyer, "Digital Pharma: Big pharma's iPhone apps (part one)," InPharm.com, July 2, 2010. http://www.inpharm.com/news/digital-pharma-big-pharma-iphone-apps-part-on
5. "About DIDGET," http://www.bayerdidget.com/About-Didget
6. "Bayer HealthCare Launches First iPhone Application with Personalized Tools to Assist People on Betaseron (interferon beta-1b) in Managing Their Multiple Sclerosis Treatment," Bayer HealthCare Pharmaceuticals, June 3, 2010, http://www.bayerus.com/News%5CNewsDetail.aspx?ID=FF2DDB50-DB0B-CDCD- D807CFE3D990D62B
7. Ozcan, "Smart technology for global access to healthcare," 2010.
8. J. M. Ruano-Lopez, *et al*, "The SmartBioPhone[trade mark sign], a point of care vision under development through two European projects: OPTOLABCARD" and *Lab on a Chip*, vol. 9, pp. 1495-1499, 2009.
9. K. Ganapathy and A. Ravindra, *mHealth: A potential tool for health care delivery in India:* Rockefeller foundation, 2008.
10. M. R. Melo, et al., "Review: Miniaturization and globalization of clinical laboratory activities," *Clinical Chemistry and Laboratory Medicine*, 2010.

Chapter Editors: Jonathan Cooper (University of Oxford), Steve Niederer (KCL)

4

A software engineering approach to computational modelling research

Publishing

The Vision

While there has been a significant investment in biological research over recent years and, over the same period, the volume of experimental data available has substantially increased, there has not been a corresponding growth in pharmacological and clinical benefits. Thus, it is important to look for improved strategies by which to extract the maximum information from existing and new laboratory and clinical datasets. The automation of data processing and analysis is essential to improve efficiency, but this must be combined with the automated evaluation of hypotheses or theories against the available data. This will provide an unbiased context for new data, and rapidly identify gaps in our understanding, limitations in the data, and potential contradictions in current hypotheses.

In the classic Popperian framework, theories must be set out in a way that makes them falsifiable, i.e. so that they can, in principle, be proven false – if my theory is that seeing a black crow makes me happy, it is hard to imagine an experiment that conclusively proves this theory wrong.

One of the reasons that the VPH initiative advocates the use of predictive models in science (computational or otherwise) is because we believe that this is a good way to set out a theory – the model provides a formal description of the theory, and what is easier to falsify than a prediction? Posing a biological hypothesis as a predictive model should thus provide an unambiguous and testable description of a proposed physiological or pathological pathway or mechanism. However, despite the fact that reproducibility is a basic requirement for science, current models are probably the least reproducible type of research output.[32] Even different implementations of what is purportedly the same model may yield significantly different results.[32]

In 2010, the VPH initiative raised this issue and promoted a petition to try to influence the editorial boards of some journals to require that models had to be shared when published. Unfortunately, the petition did not achieve a sufficient level of support. Why this was so is open to interpretation, but a good guess is that, in reality, publication policies are interlinked with how journals are operated, and how research outputs are measured through citations. With the petition, we were asking modellers to undertake additional effort compared to authors of experimental papers. An experiment can be described by the materials and methods used and, provided that reporting guidelines such as those indexed and developed by the MIBBI project[34] are followed, this is usually more or less sufficient to reproduce it elsewhere. In a modelling paper, the model is usually so complex that no textual description can provide all the details necessary to reproduce it; our suggestion would have imposed the condition that the model used should be stored in such a way that it could be reused and tested by others, which of course requires a lot of extra work above what is needed at present.

So, we pose the question – is there an alternative approach, using ICT, that can provide the benefits of automated hypothesis testing and encourage the sharing of models and data for this purpose, while reducing the burden thus imposed on modellers (and indeed scientists in general)?

By combining a proposed formalised hypothesis (a machine-readable model description) with a comprehensive list of model assumptions and links to a detailed annotated database of the experimental data, which underpin the hypothesis, it becomes possible not only to evaluate hypotheses automatically, but also to update them as new data becomes available.

To this end, we envisage a virtual environment for publishing biological hypotheses and experimental data. Biological hypotheses will be published in the form of an annotated mathematical model. Experimental data will be accompanied by a machine-readable description of the protocol followed, in a format suitable to be applied to models in simulation. Hypotheses will then be automatically evaluated against all available and relevant experimental data in the database, resulting in a scorecard of data consistent with each hypothesis. New data will be uploaded into the experimental database and will be evaluated against all relevant hypotheses.

Automating the process of hypothesis evaluation would allow experimentalists and clinicians to evaluate their findings, while providing a central repository of experimental data that greatly reduces the barriers to developing new hypotheses consistent with previous observations. Providing access to such an environment through model development and analysis tools will speed the modelling process and result in more high-quality publications, thus offering an incentive for modellers to "buy in" to the system. This framework would thus provide the tools necessary for formalising our understanding of biology, for reusing available experimental data, and for combining data and knowledge from geographically dispersed research groups; this would provide the much-needed improvements in efficiency required to justify the vast quantity and costs of high-throughput experimental and clinical data.

But assuming this scenario makes sense, what are the challenges that would be involved in realising it?

The challenge

There are numerous aspects to the proposed framework, and hence a large variety of challenges to be overcome. However, we can usefully group them into four categories. Firstly, consideration of how models and protocols may be described in such a way that *in silico* analogues of experiments may be performed. Secondly, there are many challenges surrounding access to experimental data in a usable form. Thirdly, comparison of simulation output with

experimental data is decidedly non-trivial. Finally, challenges are not restricted to technological issues – there are also sociological hurdles to overcome.

A problem that is also common to other types of VPH data and model infostructures is that of partitioning. There is always a temptation to develop a single monolithic solution, but it is in diversity that slowly competing ideas can develop. It is unquestionable that some level of homogeneity is required for the system to work, but is it possible to develop a unifying experience while, at the same time, providing researchers with the possibility of developing and deploying alternative services or alternative implementations of existing ones? Given that much experimentation will be needed before the best approach is found, the system must be developed in a way that permits alternatives for different components of the system to be investigated. This issue will be touched on throughout the following discussion.

Executing hypotheses

To enable the automated evaluation of hypotheses, they must be made available in a form that is amenable to computer analysis. The precise nature of this hypothesis encoding has a marked impact on the type of evaluation that may be carried out. For instance, if a hypothesis is represented by an executable program for a particular computer system, little flexibility is afforded – the program may only be executed, possibly with some inputs, and the outputs examined. Without additional metadata describing the context for the hypothesis and the meaning of the inputs and outputs, no useful insight may be obtained. However, this approach has the benefit of simplicity and so will be discussed further later.

We consider a hypothesis to consist of a model of a particular biological system and its behaviour under particular conditions. Using the definitions of Waltemath et al., a *model* is a mathematical representation of a biological system that can be manipulated and experimented upon, and a *simulation* is a numerical procedure performed on a model that aims to reproduce the spatial and temporal evolution (the behaviour) of the system represented by the model, under prescribed conditions. A collection of models together with a set of simulations, and potentially other procedures, thus constitutes a hypothesis.

In this context, the challenge of executing hypotheses can therefore be subdivided further:

– description of the model (or models) concerned, in a machine-readable form.
– description of the procedures to be performed, again in a machine-readable format.
– actual execution of the procedures on some computer system.

Within the broad scope of the VPH there are many different kinds of model and procedure. There is thus considerable difficulty in conceiving a system that can provide for all possibilities, though progress has been made in specific areas, notably through the use of mark-up languages such as SBML[35] and CellML[36] for model representation. It will be important to provide flexibility for expansion, while beginning with more tractable sub-problems based on these efforts. One area for research is therefore into further development of mark-up languages so that they can be used to represent all desired kinds of model. The COmputational Modeling in BIology NEtwork[37] will play a key role in overseeing the development of a set of open, interoperable and non-overlapping standards covering all the aspects of modelling in biology.

Despite the distinction drawn in the definitions above, there often remains an ill-defined line between the two concepts of model and simulation. This may be fundamental to the modelling approach, for instance in the field of executable biology, where the model is the simulation algorithm itself. When the description of biological processes builds on numerical integration, there is often a clear conceptual distinction between a model definition and its numerical simulation over space and time, but nevertheless, both concepts are sometimes merged at the level of the description formats. For our purposes, and by analogy with experiments performed on a biological system, we prefer to draw a clear distinction between the representation of a model and of the experiments (simulations or procedures) performed on it.

The description of such experiments is less well developed than that of models. However, recent work on the MIASE[38] reporting standards, and the associated SED-ML[39] mark-up language, provide a promising means of specifying what simulations are run for a given model in order to produce the results seen in published papers. The first version of SED-ML is limited in terms of the simulations that can be defined; further research will be needed to extend this to other problems.

In a recent paper[40], three of the chapter editors proposed a concept called "functional curation" and a prototype system to provide it, which aligns closely with the goals of this chapter. The basic premise is that effective reuse of a quantitative mathematical model requires not just access to curated versions of the model equations, but also an understanding of the functional capabilities of the model, and the advisable scope of its application. In other words, you need to know its behaviour when simulated. The paper is targeted at the large portion of biomedical models that can be rendered with mark-up languages such as SBML or CellML, although it considers a slightly different context. Functional curation is seen as being an extension of model repositories – as well as containing models, the extended

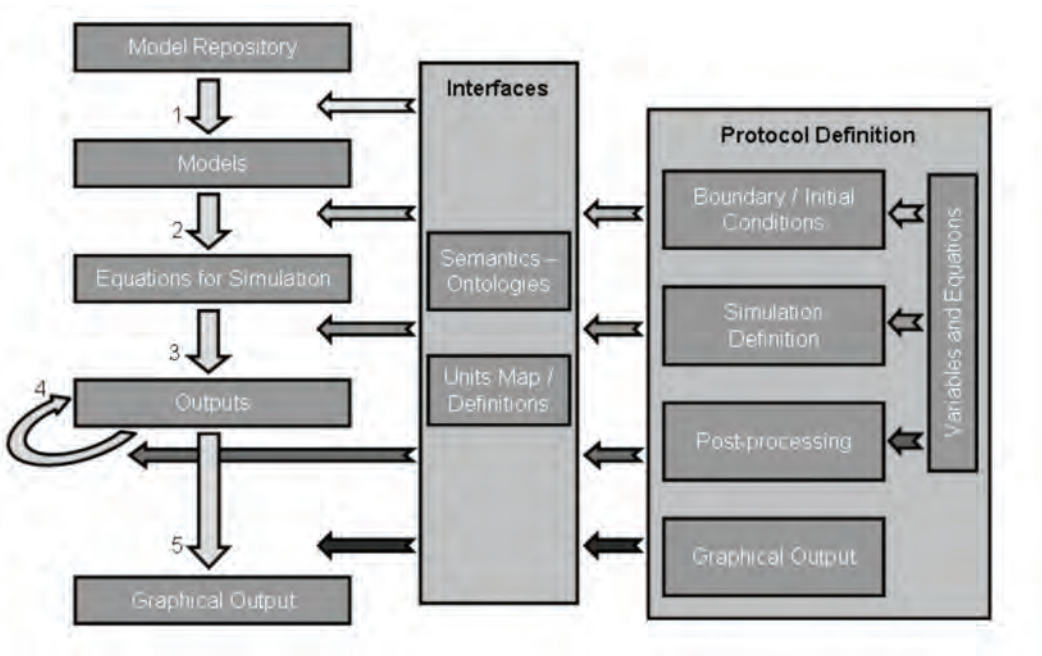


Figure 4.1: (taken from Cooper, Mirams & Niederer). Components of a functional curation system. The steps involved are: (1) annotate models with ontological tags; (2) combine model with protocol mathematics and simplify; (3) run simulation; (4) post-process model outputs; and (5) graphically plot results. All interaction between the protocol definition, model equations, and simulation implementation is filtered through an interface definition, ensuring that biophysical entities and physical units are appropriately mapped.

repositories would contain "protocols" describing the simulated version of an experiment and, ideally, reference actual experimental data for that protocol. When a new model is added that purports to represent the system to which the protocol is applied, the protocol is run on that model and the results compared (both to existing models and to experiment). Similarly, if a new protocol is added, it can be run on all relevant models. The paper describes several enhancements that would be required in SED-ML to enable it to describe protocols of the kind given.

The various components required in a functional curation system as described in the paper are illustrated in Figure 4.1. All aspects admit further investigation to a greater or lesser extent.

– **Ontological annotation of models and protocols.** Models vary in the naming conventions used for entities in the model. For a single protocol to be applicable to multiple models, some other addressing scheme is therefore needed. The natural approach is for both models and protocols to use standard ontologies for referring to biological entities or processes. There is considerable work still to be done on standardisation, but progress can be made in the interim if models and protocols in a particular sub-domain agree on a set of standard names to use. This would lead to a process of "bottom-up" ontologisation as the field develops.
– **Unit conversion.** Handling of scaling between quantities of the same dimension is straightforward, albeit not always implemented in current systems. However, more complex conversions may be needed to support the application of a protocol to models of the same system

[32] Waltemath et al., Minimum Information About a Simulation Experiment (MIASE). PLoS Comp. Biol., 2011, 7(4): e1001122. DOI: 10.1371/journal.pcbi.1001122
[33] Niederer et al., Verification of Cardiac Tissue Electrophysiology Simulators using an N-Version Benchmark. Phil Trans R Soc A, 2011.
[34] http://mibbi.org/
[35] http://sbml.org/
[36] http://www.cellml.org/
[37] COMBINE, http://co.mbine.org/
[38] http://biomodels.net/miase/
[39] http://sed-ml.org/
[40] Cooper, Mirams & Niederer, High throughput functional curation of cellular electrophysiology models. Prog Biophys Mol Biol, 2011. DOI: 10.1016/j.pbiomolbio.2011.06.003

which follow different conventions for representing the biology mathematically. For example, variables that represent the same biological structure or process may have different dimensions in the different models. One special case in cardiac electrophysiology has been addressed by Cooper et al.[41], but a more general framework would be desirable.

- **Model modification.** The functional curation approach allows protocols to change the model equations to reflect experimental procedures. For instance, a voltage clamp experiment on a single ion channel may be applied to a whole-cell electrophysiology model. Since the transmembrane potential is imposed by the experiment, the equations defining its evolution with time in the model are redundant and must be discarded. In fact, any equations that do not influence the behaviour of the ion channel in question may be discarded. Much further investigation is needed on how well this approach can transfer to models that are not ODEs, and on whether a single protocol may be applied to models using different formalisms.
- **Simulation execution.** The prototype in the paper is applicable only to systems of ODEs, and SED-ML has similar restrictions to SBML in the kinds of mathematical models that can be simulated at present. Extending the languages and implementations to other formalisms, and supporting other languages for model representation, requires further work.
- **Post-processing.** Cooper et al. suggest a small set of core operations, which can be combined to express complex protocols; however, at present, only a small sample of protocols have been investigated. The optimal balance between expressivity and ease of implementation is still unknown, and further language development will be required.
- **Graphical output.** Two- and three-dimensional plots are well characterised and understood. However different visualisations may be more effective for some kinds of model, and the representation of uncertainty is also of interest. These issues are addressed further in other chapters.

Of particular interest is the extent to which a single protocol language is applicable to different styles of models, for instance using different mathematical formulations (stochastic or deterministic, spatial or lumped parameter, etc.). This is very much an open question. Alternatively the system may benefit more by supporting a variety of protocol description languages, which introduces its own challenges in integrating the whole.

In contrast to this "white box" view of models and protocols, in which protocols possess the ability to make modifications to the model in order to express complex experiments, there are also benefits to a "black box" approach, which treats the model as an opaque entity merely possessing inputs and outputs. Implementations of a black box approach can be much simpler, being essentially just workflow engines connecting these inputs and outputs in a graph structure. Much research on workflow languages and repositories has been done, including on their application to life sciences, and this could be utilised for these simpler cases. Being conceptually simpler, workflows may also possess the benefit of being easier to comprehend by users. Whichever approach is taken, developing libraries of standard protocol components, experimental setups, input/output transformations, etc. will greatly assist new users in encoding their protocols.

Finally, whichever descriptions are used for models and protocols, we need infrastructure. This includes storage for models, protocols, and both simulated and experimental results; this aspect is addressed further in other sections. Here, we focus on the computational infrastructure to run protocols on models. We do, however, note that there are privacy implications if the system is used for running protocols not just on already-published models but also on in-development models. Clearly, it is beneficial, indeed essential, to be able to test hypotheses before publication, but care will be needed to ensure that pre-publication models, potentially containing commercially or clinically sensitive information, are not exposed to unauthorised users.

A service-based infrastructure seems most appropriate and this would allow maximum flexibility in varying the underlying computational resources as required. For many, if not most, protocols, a cloud computing resource would be the natural choice. These are now mainstream and widely available. Some protocols or complex models may require access to high performance computing resources; achieving this transparently is still, despite considerable research, often difficult. It may also be of interest to allow scientists or others to contribute their own compute resources for use, especially if specific proprietary hardware or software is required in some cases.

Whichever infrastructure is chosen, funding the ongoing availability of the service will be a crucial challenge. Various funding models are possible, including subscription-based or supported by journals. Potentially, the costs could be partially covered by funding bodies, much as open-access publication costs are at present.

Where models and/or protocols are defined in mark-up languages, there is also considerable scope for automatic optimisation techniques to improve execution times, and hence reduce the computing costs associated with the system.

Where a model or experiment consists of an arbitrary executable, security considerations exist for executing the hypothesis. The most challenging approach would be to security audit the executable (automatically) to ensure that it had no undesirable side effects or malicious behaviour. A more practical solution would be to employ some form of restricted execution, virtualised or sandboxed environment in which to run the executable. This is another argument in favour of the mark-up language approach, employing modelling languages with restricted semantics, which are therefore incapable of encoding unsafe functionality. However, some models may be implemented using proprietary software and be too specific for a mark-up language representation, or porting the models may require too much effort, making the ability to test such models still desirable.

Finally, links with benchmark challenges could also be investigated, to enable such challenges to be posed and evaluated through a system such as is proposed here.

In summary, the key issues for model execution are:

- further developing mark-up languages for model encoding.
- whether and how to support models defined purely by source code or executables?
- methods for describing "experiments" to be performed on models:
  - whether one protocol language can satisfy all use cases?
  - if not, how to support a variety of methodologies?
  - workflows for describing protocols connecting "black box" models.
  - developing libraries of standard protocol components, experimental setups, input/output transformations, etc.
- semantic annotation of models and protocols.
- computational infrastructure:
  - cloud-based model storage and execution.
  - satisfying requirements for HPC or specific systems.
  - sustainable funding model.
  - security.
  - privacy.
- support for benchmark challenges.

Data access

The key issues for model execution outlined above apply equally to data. In order to allow comparison of simulated and experimental results, the relevant data needs to be made available in formats supporting such automated comparison. This requires work both on the formats themselves and on semantic annotation to enable tools to derive meaning from, and about, resources.

Formats for storing both experimental and simulated data vary widely and are often tool specific. Furthermore, large datasets are normally stored in efficient custom binary formats making interpretation by general tools difficult. Using standard file formats, such as HDF5 or NetCDF, provides improved interchange possibilities but still requires significant effort to support. Higher-level mark-up languages (FieldML, SBRML) or annotations (BiosignalML) allow tool-agnostic description of data in a manner amenable to annotation using the same technologies as those proposed for model and protocol description.

Data on its own has little value. Value is added through annotation with information describing the source of the data, the protocol used to obtain the data, any assumptions or limitations associated with the data, etc. A mark-up language approach to storing and exchanging data provides a standard interface for the annotation of the data. Mark-up languages such as FieldML and BiosignalML provide mechanisms to link a description of the data to separate resources which actually store the data, allowing efficient large-scale data stores to be addressed and annotated in the higher level mark-up language. Proprietary data formats would act in a similar manner as black box model descriptions, and further work would be required to extract useful information, perhaps building on research into secure sharing of legacy data.[42]

As with the standard description of experiments to perform on models discussed in the previous section, one could imagine having a library of standard data descriptions, which could be used to annotate new datasets. A straightforward user interface could be provided to enable the data provider to customise a standard description to their particular dataset.

Some form of data infrastructure is also needed to provide access to data wherever it is needed. We envision an ecosystem of repositories with a common querying and access interface. The key is to provide services to make it as easy as possible for users and tools to find, access, query, interact with, reuse, share, data.

The execution of large-scale models requires the use of high performance computing resources and typically results in large scale simulated data. Wet-bench experimental protocols can also result in very large datasets. Access to such datasets is a difficult problem to address, particularly in a global sense. One possible approach is to keep large datasets in one physical location and always use local resources to extract specific subsets of data required for a particular simulation or analysis. With this approach, the large dataset does not require replication across the Internet or other means of transporting the data from one physical location to another,

but only works when it can usefully be broken into smaller subsets of data. An alternative approach is to assume that data storage and transport is an unlimited resource, and transport the entire dataset to any physical location where access to that data is required. Some kind of intermediate approach is likely to prove the most practicable, with the development of services that can interrogate a dataset to establish the best form of access for a particular purpose (copy the whole dataset, grab just the required subset, etc.).

As with model and protocol descriptions, many issues exist regarding the security, privacy and confidentially of data. This is an issue emphasised greatly in clinical data, where there is a need to protect patient identity while making the data available to the community, and much research has already been done in this area[43]. While it is reasonable to expect any computational tools developed here to work equally well with public and private data, it would seem advisable that the establishment of automated hypothesis testing is based on public data. A common scenario might be for a modeller or experimentalist to want to test existing models against their data prior to making it public in order to include the results of such testing in a publication.

Comparison of experiment and simulation

Central to the concept of automatic hypothesis evaluation is the comparison of hypothesis predictions against experimental observations. The comparison of predictions with observations is fundamental to the concept of falsifiable science and, although conceptually simple, the process of comparing experimental data with hypotheses predictions requires a rigid formalism to ensure that comparisons are valid, relevant and informative.

There are three challenges in comparing observations with predictions that are intimately linked with the sociological and data challenges described in this chapter. To ensure that the proposed framework is easy to use and general, the experimental database must be agnostic to data format. Not only does this pose challenges for storing and annotating arbitrary formats, it requires the ability to process raw experimental data into physiologically relevant quantities. Many measurements made in laboratories are recorded as voltages, currents, fluorescence intensities or frequencies, as opposed to meaningful physiological values such as ionic concentrations, capacitance, cell volume, etc. Data may also be interpreted by averaging locally (moving mean), over time (over period of a system) or across multiple preparations (different cells). Including a formal, machine readable and repeatable translation of the raw experimental measurements into their interpreted physiological values improves the transparency of results and clarifies the often overlooked link between experimental measurements, the interpretation of experimental data and the comparison between interpreted experimental data and hypothesis predictions.

Secondly, to compare a formalised hypothesis with experimental observations requires exactly the same protocol to be performed in the laboratory and within the proposed hypothesis evaluation framework. Many physiological functions are characterised by processing interpreted data through a pipeline of post-processing steps. In many experiments, a protocol is performed, measurements are made, and these are converted to a physiologically meaningful value. This process is repeated under different conditions, and a single measure that characterises the salient feature(s) of each measurement (maximum, minimum, decay rate, etc.) is plotted against the perturbed experimental parameter. In some cases, the post-processed results are the only measurements available, or by plotting results in this way a feature of a process of interest is amplified or exposed. These post-processing steps must therefore be formalised and stored in a machine-readable format that can be applied to the hypothesis predictions.

Thirdly, comparisons should only be made between consistent systems. To ensure that comparisons between predictions and observations are fair, they need to be between hypotheses developed to represent the system that produced the observations. This requires annotations describing the system (species, temperature, cell type, genetic strain, etc.) used to produce each observation.

Any quantity that could conceivably become part of a hypothesis must then be recorded. To this end, minimum information standards have begun to appear in recent years, both for description of experimental conditions (e.g. MIAME[44] – Minimum Information About a Microarray Experiment, MICEE[45] – Minimum Information about a Cardiac Electrophysiology Experiment) and for description of models and simulations (e.g. MIRIAM[46], MIASE). Encouragingly, all of these standards are working together under the MIBBI[47] (Minimum Information for Biological and Biomedical Investigations) framework.

Any hypothesis must then define the physiological scope of its predictions, for example whether it applies to a particular genetic mutation in rabbits, or whether it applies to all mammals. Similarly, quantities such as pH or osmolarity should be explicitly ‘fixed’, given ranges, or allowed to take any value. The development of minimum information standards for the experimental systems will provide a framework wherein all hypotheses must explicitly define a value/range for a quantity recorded in a minimum information database, or define the fact that the hypothesis states this quantity is irrelevant/can take any value. By so doing, we define which hypotheses apply to which experimental datasets.

This structure would greatly accelerate the inclusion of new parameters into hypotheses. For example, temperature is known to affect the dynamics of many ion channels in the heart, and it appears explicitly in some equations of cardiac

cell models, yet it is included only implicitly (with a fixed value) in the parameters of other equations. We believe that a framework such as the one suggested would drive the development of hypotheses forward, so that quantities such as temperature, species, gender, pH, etc., become explicit parameters in the hypotheses

To test a hypothesis requires a method for comparing predictions with observations. Although some observations can be posed as binary outcomes (increasing or decreasing, threshold response), the majority of comparisons to biological outcomes can be posed only as comparisons of time series data. The framework would need to provide a suite of standard evaluation methods (Root mean square (RMS), L2-norm, L-infinity norm, etc.) with a suitable default measure (e.g. L2-norm) chosen, but users would have the ability to define mathematical functions for comparing data, thus providing an extendible and customisable tool for data comparison. This would provide a mechanism for evaluating the goodness of fit for predictions to observations.

To evaluate the quality of data, the framework would exploit the consistent format used for describing predictions and observations; this would allow both predictions and observations to be compared against all consistent observations and hypothesis predictions. In this way, the converse of prediction checking could be used to evaluate how well one observation matches all hypothesis predictions, and as well as other observations to provide an initial and naive method for ranking observations. More weight would be given to hypotheses that are validated or contradicted by higher ranked observations, leading to a default method for ranking observations, though customisable weighting would be essential to counter the effect of prolific publications of poor results or to give more weight to new and improved measurement techniques.

To convey the results, two formats would have to be introduced. The first should focus on an individual hypothesis or observation and could summarise in a table the results of comparisons with all other predictions and observations. The second option should provide users with the ability to define graphical outputs tailored to their data; this would require the creation or adoption of a graphical output mark-up language linked to observations. This would be an essential tool for model verification and would greatly facilitate the model review process.

User Interface

A monolithic software solution to our vision is unlikely to gain general consent. Yet, for our vision to have any chance of success, we need critical mass.

A step in that direction is to opt for an open source approach to software development. Several web-based hosting solutions exist (e.g. GitHub, Google Code, SourceForge), supporting different version control systems, be they distributed (e.g. Git, Mercurial) or not (e.g. Subversion, CVS). An open source project is usually distributed under one or several specific licenses. The choice of such a license can be tedious, but is essential since some licenses may unnecessarily prevent use of the software in certain cases[48].

Any software solution developed as part of our vision should, whenever applicable, offer support for the three main operating systems used by our community (i.e. Windows, Linux, and Mac OS X). It should also rely on agreed standards (such as SBML and CellML) if it is to ensure consistent access to data, protocols and models. To avoid duplication of work and to reduce the risk of solutions of sub-standard quality, well tested, documented and supported APIs for these standards should be used. If no such API exists, then it would be in the interest of the community to provide an implementation for it, which raises the potential issue of instigating community involvement. This can generally be facilitated by having well-documented code, and by providing clear instructions on how to develop, build, test and package such code.

Whenever possible, reusable components should be used or developed. For example, a cardiac action potential analyser, which can extract key parameters from a signal, is likely to be relevant to several software solution, so it should be developed from the outset with the view to being self-contained. This would allow other developers to provide their users with its features; ideally, this would be achieved through a plugin mechanism. Unfortunately, software solutions tend to use different plugin interfaces, but as long as a feature relies on self-contained components, its conversion into a plugin ought to be relatively straightforward. Most importantly, such an approach would give the user control over the software.

Indeed, no matter how well engineered a software solution is, its success (or failure) will be determined by its users (or lack thereof). The software must therefore have a low barrier to entry and yet be flexible (e.g. work both from the command line and through a graphical user interface, be fully customisable), as well as powerful (e.g. offer additional features through the use of plugins), so as not to discourage advanced users from using it. The lack of adequate documentation, if any exists at all, has often been the downfall of otherwise promising software solutions. It is therefore essential to spend the time and resources to provide users with suitable support, if we are to gain their loyalty. Interaction with users is also important, be it through a mailing list, a forum and/or some other means.

Most importantly, however, a software solution must provide its users with added value. For example, our vision will provide researchers with a wide range of data, protocols and models, all of which will be readily accessible online and/or through software, providing researchers with information which will help them to produce more relevant models,

protocols or even data which could, for example, result in an increased chance of getting published in a high-impact journal. Further tools could also be developed to leverage the services provided, for example by allowing modellers to fit their models to a wider range of experimental data under multiple protocols.

Added value would also be provided through support of both SBML and CellML, the two main mark-up languages commonly used to describe biological models, and their conversion from one to another. Support for other formats (e.g. JSim's MML format[49]) could also be added through an import/export mechanism. Access to the CellML Model Repository[50] and the BioModels Database[51] through web services would also help end-users with, for example, finding and downloading models.

Besides these software considerations, there is the issue of sharing models, data and protocols. Some of this information may be found in public repositories, and may even be available in a relevant format. Other aspects may be obtained from the literature, but would then have to be converted, digitised, etc. and/or the authors contacted to gain access to the original information. In some cases, the information may be known, but unpublished, while in others such information is not to be made public before some time, if at all (e.g. a group may want to keep its technical edge, or the study is not yet complete). There are, however, plenty of opportunities for those willing to widen their research interests, start new collaborations, etc.

To obtain an initial critical mass of data, it seems likely that more progress will be made by encouraging modellers to contribute the data used in developing their models, especially if the tools provided to support such development interface directly with repositories connected to the system. The incentives for experimental groups to contribute are less clear, though some benefits can still be anticipated (e.g. provision of facilities for long-term data storage, ease of querying, etc.; the system would also assist those interested in checking if existing models could be of use in interpreting their data).

The VPH motivation

The essential motivation for this topic has already been given in the Vision section. The proposed developments will lead to a culture shift in computational modelling research within the VPH, providing the same type of epistemological solidity that is proper of experimental methods. Together they form a collective framework for the presentation of detailed hypotheses and the observations on which they are based and against which they must be tested, along with the infrastructure to enable their robust analysis.

The framework described will also support reliable model reuse, which is vital for progress as models continue to develop in complexity. It would enable the identification of models that exhibit particular experimentally observed phenomena and would also be used to ensure that any incremental development of models to investigate new scenarios does not remove their ability to replicate desirable experimental results.

Impact on Biomedicine

The biomedical literature is full of clear examples: until the modelling speculation is confronted with enough observations to build a reliable falsification framework, the translational relevance of that modelling research for biomedicine is negligible. The current problem is that, in most cases, complex models can be falsified only if those who develop the model can also develop the falsification experiments and then publish a paper that includes both activities. This, of course, greatly limits the opportunities, as in some fields only very few groups possess both of these skills.

If an editorial model for computational modelling research such as that described above were available, we are convinced that a much larger number of computational models could be brought to a level of validation sufficient to start making an impact in biomedical research and clinical practice.

General Impact

The problem here described is equally important for all those computational sciences that aim to translate their models to an empirical context: engineering, biomedicine, applied chemistry, environmental research, etc. Thus, the impact potential outside of biomedicine seems considerable.

Editorial Team

Name	Affiliation
Alan Garny	University of Oxford
Molly Maleckar	Simula Norway
Gary Mirams	University of Oxford
David Nickerson	University of Auckland
Marco Viceconti	Istituto Ortopedico Rizzoli

[41] Cooper et al., Considerations for the use of cellular electrophysiology models within cardiac tissue simulations. Prog Biophys Mol Biol, 2011. DOI: 10.1016/j.pbiomolbio.2011.06.002

[42] Russell et al., On the secure sharing of legacy data. In Proceedings Information Technology; New Generations, 2009. DOI: 10.1109/ITNG.2009.21

[43] See also the VPH ToolKit data guidelines at http://toolkit.vph-noe.eu/home/methods/toolkit-guidelines/data-characterisation-guideline.html

[44] http://www.mged.org/Workgroups/MIAME/miame.html

[45] Quinn et al., Minimum information about a cardiac electrophysiology experiment (MICEE): Standardised reporting for model reproducibility, interoperability, and data sharing. Prog Biophys Mol Biol, 2011. DOI: 10.1016/j.pbiomolbio.2011.07.001 – http://www.micee.org/

[46] http://biomodels.net/miriam/

[47] http://www.mibbi.org/

[48] For example, GPL v2 and v3 are not business friendly, see http://toolkit.vph-noe.eu/home/methods/toolkit-guidelines/licensing-guideline.html for more information

[49] http://www.physiome.org/jsim/docs/MML_Intro.html

[50] http://models.cellml.org

[51] http://www.ebi.ac.uk/biomodels

Chapter Editors: Gabor Szekely, Andreas Schuppert

5

Disease modelling and virtual physiology

Diseaseome

The Vision

The past decade has witnessed tremendous progress in the creation of powerful tools for the simulation and modelling of living systems. Today, systems biology offers numerous approaches capable of quantitatively characterising cellular and molecular regulatory processes in unprecedented detail. Systematic efforts in the field of VPH research have resulted in complex, often quite comprehensive, mechanistic models of human physiology at the tissue and organ level.

Unfortunately, the transition between a physiological and a pathological state is still poorly modelled in general terms, in spite of its vital importance to all clinical applications as well as to drug research. This is mostly caused by the inherent complexity of disease development, which is a multi-factorial process which can only be understood by accounting for the interplay of all related aspects, like genetic and epigenetic factors, the physiological state of the effected organism and the influence of numerous environmental parameters.

The fundamental long-term target of the research envisioned here is the development of essential components of a generic framework that will enable the comprehensive characterisation of the set of diseases, the diseaseome, in its full complexity and heterogeneity, by allowing the integration of the modelling and simulation approaches available today in all aspects mentioned above. However, to retain a scope suitable for the VPH-FET roadmap, we shall restrict our vision to the impact of diseases on the human physiology and its consequences for clinical applications. This will allow the full utilisation of the current state of, and future progress resulting from, the ongoing investigations on molecular, genetic or pathogen-driven causes of the diseases that are taking place in numerous public and privately funded research programmes, at both an EU and a national level.

The challenge

In general, diseases can be represented by variations of physiological parameters that lead to a reduction of the functionality of organs or their mutual control, both accompanied by a significant negative impact on the functionality of the overall system, in extreme cases with fatal consequences. The reduction can have direct or indirect consequences.

- The decreased performance of an organ in normal life, such as chronic obstructive pulmonary disease (COPD), which leads to reduced capacity of the lung for oxygen uptake and results in a significant reduction in the quality of life.
- A significant reduction of the range of homeostasis, which leads to a reduced stress tolerance of the system, such as allergies but has no impact on stress-free life. In extreme cases, some diseases based on genetic disorders, such as Gilbert's Syndrome, have no clinically relevant impact. However, they reduce physiologically relevant parameters, such as the excretion or metabolisation of drugs and so lead to toxic adverse drug effects for otherwise uncritical drugs or administration regimes. As these disorders do not reduce the quality of life in the unstressed state, their classification as disease may be unclear, although neglect of the disorders may lead to unforeseen life-threatening effects in stressed situations.
- An increase in the risk of fatal system failure without a visible impact on system performance (e.g. hypertension increases the risk of stroke but does not directly effect the performance of the system).

Many well known diseases, such as cancer, infections or neurodegenerative disorders like Alzheimer's disease, lead to severe consequences for life, such that there is common sense in characterising these disorders as disease. Other diseases, such as ADHS, lead to impacts on life that depend on cultural conditions, while others, such as increased cholesterol levels, have unclear consequences on the overall lifespan or quality of life.

Because of the tremendous complexity of the diseasome and its interaction with physiology, modelling offers a multitude of challenges.

a) Modelling the interface between –omics-based disease models and mechanistic models of human physiology

- Genetic/proteomic/metabolic aberrations from the "healthy" status affect physiological parameters resulting in a clinically abnormal disease state. Disease modelling has to provide the quantitative interface between the physiological parameter set p describing the virtual physiology at a macroscopic level and the set of abnormal –omics parameters. For example., a genetic mutation can result in an over-expression/suppression of protein levels or in reduced enzymatic activity of proteins which may be compensated by changed expressions of other proteins or result in a shift of a physiological parameter. If the parameter exceeds a threshold value, then a disease may develop at a clinical level. Hence, the sensitivities of the physiological model parameters p on –omics parameters have to be quantified with good accuracy in the range of the thresholds.
- Models quantifying the impact of a disease-related shift at the –omics level on the human physiology at the macroscopic level across cell–tissue–organ levels are intrinsically multi-scale models across a multitude of scales. Hence, efficient multi-scale modelling approaches which are adapted to the special structures of scale interactions in biological systems are required.
- Special emphasis must be placed on heterogeneous time scales, which are relevant in the dynamics of disease

development and the response to therapies. Processes across a multitude of time scales may be relevant for disease development and therapeutic effects, ranging from seconds and minutes for phosphorylation levels to months and years for shifts of clinically relevant parameters in chronic diseases. Moreover, non-linear dynamics plays a significant role, resulting in dynamic features such as phase locking, time lags or hysteresis. Reliable disease models have to reflect these dynamic phenomena in an efficient mathematical formulation.

– An appropriate definition of homeostasis in terms of physiological model parameters will be a crucial challenge for disease modelling. As mechanistic models based on a quantitative biomedical understanding of the underlying mechanisms will rarely be available, there is a need for data analysis and parameter identification tools that will support an effective characterisation of the homeostatic range for a physiological model allowing to separate sub-clinical from clinically relevant physiological states.
– Special attention should be paid to the impact of genetic variation of populations on disease development. Genetic variation can change the sensitivity of physiological parameters with respect to pathogens, toxins or drugs, which leads to a significant impact on health care workflows. Hence, there is a need for a modelling framework that supports the simulation and analysis of the effects of population heterogeneity on diseases using established disease models. Most straightforward might be a Monte Carlo sampling of –omics parameter combinations resulting in a distribution of disease models instead of a single model. This distribution could be used as an underlying prior for Bayesian modelling approaches.

b) Modelling the effects of co-morbidities resulting from joint genetic background of diseases

Human diseases result from abnormalities in an extremely complex system of molecular processes. In these processes, virtually no molecular entity acts in isolation, and complexity is produced by the vast amount of dependencies between molecular and phenotypological features. In large-scale meta-analysis (e.g. Schadt et al., Nature Genetics, doi:10.1038/ng1589), the mutual involvement of genes and diseases have been analysed at a genome-wide level. It appears that one-to one relations between genes and diseases are exceptions, restricted to the Mendelian diseases – complex diseases such as cancers, metabolic disorders, autoimmune diseases or psychotic disorders, are not associated with only single genes.

Although associations between the genotype and physiology may neglect relevant biological mechanisms such as epigenetic control, transcriptomics, protein phosphorylation etc., the studies showed that complex diseases are, in any case, associated with large sets of genes. Moreover, the relations between diseases, or groups of diseases, and the respective gene sets are not one-to-one. Most of the disease-related gene sets showed significant overlaps indicating relationships of complex diseases at the genome level. Using these overlaps, both at the level of the diseases and at the genome level, a relationship network can be established in which either two diseases are connected if the respective gene sets show a significant overlap, or two genes are connected if they are related to the same disease. However, the resulting networks showed a very high degree of connectivity so that it has proved difficult to extract clinically relevant conclusions from networks connecting only genotype and clinical phenotype. Moreover recent results show that GWAS studies linking complex diseases and genotype could not lead to results of therapeutic relevance [J. Couzin-Frankel, Science, DOI: 10.1126/science.328.5983.1220]. Hence, more advanced multi-level network approaches linking genomics, proteomics and metabolomics with clinical phenotypes may be required.

In particular, it has been stressed that combining genomic, proteomic, metabolomic and environmental factors may provide insights into pathogenomic mechanisms and lead to novel therapeutic targets. For example, psychiatric disorders, in particular, seem to lend themselves to systems-based analysis. It is well known that schizophrenia has a strong genetic compound with concordance rates in monozygotic twins reaching approximately 50%. This increased risk is conferred by a multitude of different genes, with the most important genetic polymorphisms accounting for only 1% of increased risk. It seems likely that the disease is ultimately precipitated by a complex interplay of genetic predisposition and a broad spectrum of environmental and nutritional factors. In this context, epidemiological factors such as urbanicity, geographical distribution and migration behaviour, and maternal risk factors such as infections, malnutrition and adverse life events during pregnancy have been suggested as being associated with the risk of schizophrenia onset. The relationship between these factors and the interplay with genetic determinants remains unknown and integrated, system-based investigations are a promising approach to obtaining deeper insights into the disease aetiology.

The network-based description of complex diseases can be structured in the form of network layers. The top layer is formed by human beings and the connections between them, such as family relationships. The second layer is generated by human diseases. Notably, the diseases are linked to each other as many illnesses have related pathologies or even causal relationships. An intricate example is the increased prevalence for diabetes in schizophrenia patients (approximately 15%), which seems

to be directed as diabetes patients have not been reported to have an increased prevalence of schizophrenia. Known links between diabetes, obesity and asthma have been brought into context with the representation of disease commonalities in networks. Interestingly, an increased risk of obesity can also be found in schizophrenia patients and the resulting cardiovascular diseases mainly account for their higher mortality rate. The third and bottom layer is formed by molecular systems, and the network represents interactions between these molecules such as involvement in common pathways, protein-protein interaction or co-expression.

The different layers of networks are highly inter-dependent. The network of patients and the network of molecular interactions are closely linked depending upon which molecular abnormality is present in which patient; this information can be represented in a directed graph. As the structure of the graph depends on the disease under investigation, the properties of the graph may give important clues about the aetiology of the disease. This network concept has been applied in the context of human cancers, and it was shown that networks can be used to infer biologically relevant dependencies between risk-conferring genes.

The structure of the links in the network layers can give insights into disease complexity. At the molecular level, disease-related abnormalities can arise from single biochemical systems as reflected in alterations of only a few molecules. In multi-genic disorders, this is normally not the case, and the high complexity is reflected in abnormalities in a large number of genes, molecules and the accompanying clinical symptoms. In psychiatric diseases, the clinical features of different patients are rarely very similar; this lack of homogeneity is directly reflected in the patchiness of the graph linking patients and disease-related abnormalities. Schizophrenia, for example, has a broad spectrum of clinical manifestations, which resulted in the hypothesis of the existence of diverse underlying etiologies. It is essential to be aware of this degree of patient heterogeneity to be able to determine disease-intrinsic molecular abnormalities. The known relationships between different diseases are ultimately reflected in similarities of the underlying molecular pathologies, so the network of disease relationships is intrinsically linked with the network of molecular functions. The disease network could thus be seen as a collection of disease-specific clusters that interact with each other, depending upon how related the biochemical underpinnings of the different diseases are. Information from different diseases and commonalities as well as differences of the molecular information may thus give important leads about pathological mechanisms, diagnostic applications and novel drug targets.

c) Modelling of disease development on physiological level

Modelling of disease development requires quantitative dynamic models to be established for the shift of the physiological parameter set p as a function of time: $p = p(t)$. This function can either be modelled explicitly or as a function of the root cause of the disease. For example, the root cause of the disease may be a tumour, such that the growth of the tumour size s affects a physiological parameter p of an organ. Then the time course of p can either be identified as a function $p=p(t)$ or as a function $p=p^*(s(t))$. In the latter case, the time course of the physiological parameter is modelled via nesting of the impact of the tumour size s on the physiologically relevant parameter p and the time course of s. This might be beneficial if both functions can be established independently, e.g. if reliable growth models of the tumour exist and if the quantitative effect of tumour sizes on the physiology parameters p is established independently by the function $p^*(s)$. In practice, even a single parameter p may depend on multiple tumour parameters, so that the identification of p^* may require estimators for multivariate nonlinear functions. In the case of a high complexity of p^*, it might be beneficial to represent p directly in terms of an empirical function p(t) which has to be learned from data.

A special challenge in modelling disease propagation will be the optional discontinuity of the function p(t). Even if the root causes of the disease at the physiological level, e.g. the tumour size or the viral load in a viral infection, are smooth and monotonic functions in time, the effect on the physiology, expressed by the parameter p, may become discontinuous if the function $p^*(s)$ shows discontinuities, e.g. the representation of organ failure due to tumours exceeding a critical size. Because of the high redundancy of biological functionality and its intrinsic robustness against failure, a reliable characterisation of these critical parameters may require very exact models at a detailed level. Detailed studies on networks have shown that random deletion of up to 70% of the nodes in a network does not necessarily affect the overall functionality of the network [e.g. Stelling et al., Nature 420, 190-193 (14 November 2002) | doi:10.1038/nature01166]. However, targeted deletion of a very few "hub" nodes can result in fatal network failure. As detailed modelling can result in very high costs for data acquisition and validation of the models, efficient shortcuts using reduced models have to be developed which allow reliable estimates for critical parameter bounds which are not too far from the true values.

d) Model based identification of Biomarkers for early identification of disease

Another challenge in modelling disease propagation is the identification of biomarkers for the detection of early phases of disease development. The identification of early disease states is crucial for any preventive approach in health care

and could have a tremendous impact on healthcare systems. However, while identification of fully developed diseases can be performed by classification of well separated physiological or –omics parameter sets, a clear separation of the parameters between healthy and early disease states cannot be expected. Hence, early identification requires a reliable characterisation of the bounds of the physiological parameter range, which is associated with the "healthy" state, but because of the complexity of the "critical surface" in the multidimensional physiological parameter space, a reliable characterisation may not be feasible for complex diseases.

Alternatively the identification of specific signals that indicate the crossing of critical parameters may be sufficient. As the identification of signals for criticality does not require the a priori characterisation of a critical surface in the physiological parameter space, the identification of generic signals for criticality may provide a feasible workaround. This approach may be based on the established models for phase transitions in statistical thermodynamics, which identify generic signals for criticality, based on specific features of the systems noise.

e) Quantitative modelling of the dynamics of disease development and its interplay with physiological response and therapy

Many diseases, like infections, are characterised by external pathogens which are the root cause of the disease. The pathogen will become clinically relevant if the intrinsic physiological control systems, e.g. the immune system, cannot compensate the effects caused by the pathogen. Hence a special challenge for modelling this type of disease will be a quantitative, reliable modelling of the dynamics of pathogen growth, optional changes of the pathogen population under immune/drug stress and the resulting dynamics of the immune response. Paying attention to autoimmune diseases could offer an additional pathway towards a better understanding of this transition from disorder to disease and to the identification of potential failure mechanisms in the underlying regulatory dynamics.

f) Modelling of diseases based on loss of control

These diseases are characterised by either genetic or epigenetic changes at the –omics level which result in a critical shift of physiological parameters. Due to positive feedback loops, which aim to provide rapid compensation of unexpected shifts at the physiological level, the disease-induced shifts of the parameters at the physiological level may result in a continuous, accelerated increase of the underlying changes at the –omics level. Then, normally beneficial positive feedback loops lead to lack of control if specific disorders occur at the –omics level. Examples of relevant disease types of this class are:

– inflammations, characterised by overshooting of immune response on pathogens
– allergies, characterised by response of the immune system on non-pathogens
– autoimmune diseases, characterised by a loss recognition of own cells by the immune system
– malignancies, characterised by loss of proliferation control of cells, accompanied by efficient mimicry of the malignant cells to hide from immune attack.

A challenge for modelling of these diseases is a reliable quantification of the respective feedback loops that are involved in a regulatory system which is out of control.

The VPH motivation

The major motivation for this type of research from a VPH perspective is to avoid the treatment of physiology and pathology models as two separate domains. While this might be acceptable in the classic reductionist approach, the VPH integrative approach requires this partitioning to be overcome. At the same time, we expect major improvements in understanding of the dynamic interplay between healthy physiology, pathogens and therapeutic actions. Developing tools capable of accounting for both the discrete and continuous nature of the transition between the healthy and diseased states will open a fundamentally new understanding of disease formation and, ultimately, will allow us to revise the current, rather phenomenological, ontology of the diseasome and derive some general principles from the more systematic and principled view that will be created.

Impact on Biomedicine

The fundamental motivation for the envisioned framework is biomedical: defining the transition to the diseased state means diagnosis; the dynamics of this transition and the phenotype transformations associated with it means prognosis; the external interaction with the system in the attempt to bring it back to the un-diseased state, or to a diseased state in which the associated phenotype is less disabling for the patient, is treatment. Diagnosis, prognosis, treatment: by modelling the transition to the diseased state, we can explore the foundational aspects of biomedical reasoning and open up possibilities of huge numbers of technological applications in virtually all branches of biomedicine.

General Impact

The human body is the most complex autonomous system ever created, making it the ultimate challenge for every modelling and simulation effort. Its functioning principles provide an adaptability that continually supports appropriate reactions to external threats, such as pathogens, which has always been an inexhaustible source for inspiration in science and engineering. A lot of what will be learned during the research envisioned above can be used to recognise the insurgence of, predict the effects produced by, and plan the recovery from dysfunctional conditions in complex systems.

Such knowledge can be applied, for example, to the development of realistic models for macro-economies, to ecosystems, to transportation or to information management, to name but a few. If we can develop modelling strategies that can represent the dynamics of dysfunctional conditions in the human body, there is a very good chance that these methods can also be successfully applied to less complex dynamic systems.

Chapter Editors: Feng Dong, Rod Hose

6

Visual analytics for exploratory analysis of complex results sets

Analysing

The Vision

There is a frequent need to extract classification rules for clinical decisions from complex simulation results sets. In some cases, the use of metrics can be potentially misleading, as they tend to amplify differences between scenarios that appear quite similar to the human eye. Is it possible to use visual analytics as an exploratory tool to compare different indicators over large databases of simulations? Can we use it to highlight differences between similar simulation results? Can this be unified with the assisted annotation of large collections?

Visual Analytics "combines automated analysis techniques with interactive visualisations for an effective understanding, reasoning and decision making on the basis of very large and complex data set", quoted from the roadmap outlined by the VisMaster project[1, 2]. Figure 6.1 provides a brief illustration for a visual analytics process. As a major multidisciplinary field, visual analytics includes the science of analytical reasoning, visual representations and interaction, data representations and transformations, production, presentation and dissemination. With data increasing in volume and datasets increasing in size, visual analytics has been receiving an increasing level of attention.

Visual analytics was mainly initiated by the unprecedented data volume in our society. A major challenge of the 21st Century is the "Information Big Bang". Over the period 2002 to 2009, the amount of Internet data increased by a factor of 56, from 5 exabytes to 281 exabytes; 210 billion emails and 4 billion SMS messages are sent on a daily basis; 48 hours of new videos are uploaded to YouTube every minute; multi-camera networks incorporating tens to thousands of cameras are generating huge amounts of data every day, more than can ever be inspected manually.

This data explosion is reflected also in science and technology, where the current trend of rapid data increase is likely to continue well into the future owing to the continued advances in scientific measurement devices. In biology, scientists are facing three billion base pairs of genes per human, while in medicine, the growing resolution and new modality of medical images has resulted in a rapid increase of medical data from gigabytes to terabytes.

With the arrival of the new data come many opportunities, including new knowledge discovery arising from the improved spatial and temporal precision. For example, advanced modelling in VPH relied on the collection of high quality data, as in LHDL [3]. Often, it is of great importance that the new data is adequately processed in good time, but the size and complexity of the large datasets can significantly increase the level of challenge and demands on the data processing techniques. The complexity arises not only because of the size of the data but also because it often comes from heterogeneous sources and is presented in dynamic, irregular, ambiguous and sometimes even conflicting form.

While significant progress has been made in highly automated data analysis, we are still far from being able to process data with full automation. To this end, new solutions are urgently needed to cope with the vast data volume and to allow for timely retrieval of valuable information. To compensate for the limited capability of automatic techniques, human participation becomes vital. Visual analytics is designed to promote the integration of expert knowledge with the data analysis process through interactive visualisation.

Visual analytics is NOT simply putting old techniques together. The collaboration between the multiple disciplines leads to novel, highly effective analysis tools, contributing solutions to the information overload problem in many

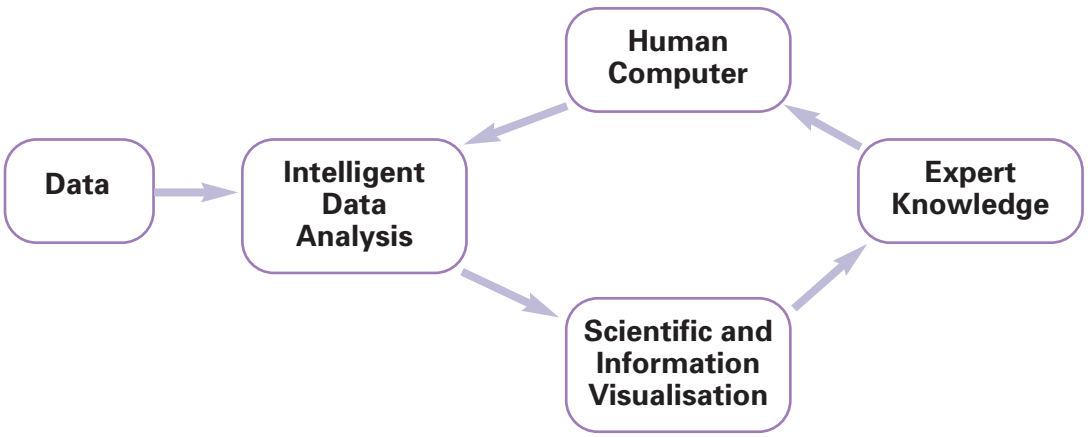


Figure 6.1: A brief illustration of Visual Analytics Process: new knowledge is discovered by the integration of intelligent data analysis with visualisation and human computer interaction.

important domains. Integration of these disciplines into visual analytics will result in a set of well-established and agreed-upon concepts and theories, allowing any scientific breakthrough in a single discipline to have a potential impact on the whole visual analytics field. In return, combining and upgrading these multiple technologies onto a new general level will have a great impact on a large number of application domains.

Current research in visual analytics is still, to some extent, in its infancy – while there have certainly been very advanced solutions for very specific examples, there is no general breakthrough yet. Automatic data analysis has achieved great success in the last few decades. Visualisation is effective in terms of helping domain experts to understand data by offering crucial visual information. However, due to the rapid evolution of data complexity, existing simulation and fully automatic data analysis techniques are often not able to reach an optimal solution. It is found that a more effective approach can be to involve the power and versatility of human decision making by integrating simulation and data analysis with interactive visualisation. While there has been some progress in this direction, much of the work is still not able to fully explore the power of integrating expert knowledge.

The effectiveness of visualisation for clinical applications has already been demonstrated through many VPH projects, such as VPH2 [4], VPHOP [5], etc. A distinctive feature of VPH is the handling of large, complex and multiscale human data for clinical decision making. Often, the complexity of the clinical data and the diversity of the patient population imply that the simulation and data analysis results obtained can serve only as a reference to assist clinicians in making a final decision. To allow for effective involvement from the clinicians, providing a faithful visual representation of the data and a good user interaction is important. To this end, visual analytics fits well into the demands of exploratory analysis within VPH.

The challenge

The critical challenges of VPH come from the exponential increase of data and the level of complexity involved within the clinical cases. To become an effective tool for VPH, visual analytics has to address the following challenges:

1) Management of models in VPH.

At the heart of VPH there are integrative models. These are models of models, models formed by component models and relation models. Component models are reductionist models that describe the phenomenon at one characteristic spacetime scale; relation models define how the prediction of our component model relates to the inputs of another, defining the systemic concatenation.

If we see an integrative model as a black box, it would have a few inputs and a few outputs, and conventional tools in most cases would be adequate. However, in reality, when we develop an integrative model, the focus is mostly inside the box, and in many cases we need exploratory tools that show how the primary outputs of the integrative models are related to all of the inputs and outputs of all of the component and relation models.

Visual analytics needs to operate on an "open box" strategy. The black box strategy shortcuts the understanding of the underlying technical principles in depth and hence suffers many limitations. Clinicians and VPH domain experts rarely have a strong technical background or deep technological understanding, so they are often confused by choices among many algorithms. Although technical comparisons of some algorithms are available, they are usually conducted in the limited context of applications and are sometimes even based on in-house (rather than real-world) data. Often, the clinicians and VPH domain experts find that methods proposed in the scientific literature work far less well in a real application than in the laboratory. Due to a lack of guidance, they have to proceed with an ad hoc trial-and-error methodology, trying different methods based on guesswork, which is ineffective and time consuming. Further, a significant level of technical adaptation and extension of the standard techniques is usually essential in order to achieve the desired outcomes for a specific application. This is not achievable with the black box strategy.

In addition, many algorithms rely greatly on the configuration of several parameters, which may have a significant impact on performance speed, accuracy, and robustness. Tuning these parameters should be performed by an expert with knowledge of the underlying algorithms. Principles for tuning models for specific data are not available, so testing model parameters remains difficult.

2) Visually controlled data mining for VPH to support the sound integration of user interaction for effective data mining.

Due to the complexity of VPH data and simulation results, it is difficult to find optimal solutions automatically in many clinical cases. To achieve a good solution, human knowledge is extremely important and this is provided manually by the clinicians. The research question is how to facilitate the human interaction and how to reach a good balance between the level of automation and human intervention.

A visual analytics system cannot be too demanding on the input (i.e. time & labour) of the clinicians, but while a higher level of automation will help to reduce their workload, it may also lead to unreliable results. In developing such a system, it remains vital to allow expert users to retain control and decide whether to accept or reject a solution.

It is particularly important to support the handling and visualisation of uncertainty information during the visually

controlled data mining. Uncertainty is a very common phenomenon in VPH and may arise from a number of sources, including data noise, technical limitations, etc. Most data analysis techniques give approximated rather than accurate measures – they create a considerable amount of uncertainty, which may significantly influence the result. Moreover, in an application that involves a number of cascaded steps, the uncertainty is likely to be propagated and aggregated at each step and ultimately result in a non-ignorable impact to the overall outcomes, which is hard to foresee without proper reasoning and visualisation. Thus, uncertainty should be taken into account in the decision making of the clinicians and VPH domain experts with respect to the selection of algorithms and parameters. In particular, when supporting incremental analysis of large-scale data, the uncertainty modelling and analysis should be performed in a hierarchical (multiresolution) form.

Visualisation techniques that are well suited to the exploration of data and model relationships within VPH include the interactive visualisation of very large models and data, multiscale visualisation (in both spatial and temporal domains), visualisation of uncertainty within models and data, and collaborative visualisation.

The visualisation of high dimensional datasets has been of research interest for many years in information visualisation. However, while there are numerous cases of specific success stories, the more general problem still remains largely unsolved, and a convincing way to represent high dimensional information is still lacking.

The visualisation of multiscale data set has achieved great success in many areas such as Google Body [6], human neural systems, etc. However, such a demand also exists at multiple scales in the biomedical area, and suitable techniques have not been fully elaborated. In fact, research is still at an early experimental stage [7].

To further open up the technical black boxes, we could also consider the use of interactive algorithm visualisations, which dynamically depict the data structure and the action of the algorithms using computer graphics and animations. There are already many algorithm visualisation resources available, though mainly for educational purposes [8]. Visual analytics should assist clinicians and domain scientists in technical comprehension by making use of existing algorithm visualisation resources and developing interactive algorithm visualisation methods. The interactive algorithm visualisation should show not only the process of the algorithms, but should also allow user intervention and adjustment.

3) **New interaction techniques to support effective model/data exploration through user interaction, e.g. eye tracking, human gesture, touch screen etc.**

Special attention should be paid to interaction techniques for collaborative and distributed working environments.

The limitations of the classic human computer interaction through the mouse and keyboard has already been generally recognised. In computer graphics and computer games, many computer vision techniques are adopted to facilitate more "natural" ways of human computer interaction by tracking human gestures and eyes, which can be adapted in VPH for clinical applications. Noticeably, multitouch techniques, which support intuitive ways of human computer interaction, have attracted a lot of research and commercial interests in recent days, and the topic can be further explored in the clinical context within VPH.

Given the scale of the VPH data set and high dimensionality in the parameter domain, intelligent data analysis is often time consuming. Progressive and steerable computing provides an attractive option by firstly working out quick answers before making incremental improvements. User interaction can be engaged in the progressive computing processing to allow for steerable analysis towards areas of interest. Further, many data sets are dynamically updated nowadays due to their fast evolution. A dynamic data computing process is needed, which requires the capability of incremental analysis and learning from the data source instead of restarting the whole process from the beginning at each time of update.

The interactive computing needs to be supported by new computing paradigms such as grid computing and Graphics Processing Unit (GPU) to gain greater computing power. Special attention should be paid to the use of the GPU which continues to achieve an exponential level of speed increase compared to the evolution speed of the CPU. Another advantage of the GPU is that it is available on almost any platform, including PC, laptop, mobile device, etc. Its local computing power makes it easily accessible, so it can help clinicians and domain scientists work efficiently while avoiding the many legal and ethical issues associated with remote computing.

4) **Evaluation techniques to test the results of the above techniques.**

Evaluation is the key to clinical applications. To prove the effectiveness of visual analytics in VPH, adequate evaluation will have to be carried out by users (i.e. doctors). The motivation of the evaluation would be to provide evidence of any improvements that a visual analytics system has brought – more specifically, how visualisation techniques have helped to highlight data information which was not clearly visible in its original form; why simulation and automatic data analysis failed to achieve optimal solutions in the cases

evaluated; and how the iterative process of visualisation, simulation, automatic data analysis and user interaction in visual analytics has helped to achieve better results.

The VPH motivation

Combining VPH and visual analytics will bring mutual benefits. "Visual analytics" became a recognised term not very long ago. However, similar types of practice have been carried out over many years in many applications of exploratory data analysis, visual data mining and visual data exploration. The recent interest in visual analytics further confirms the future trend of the close collaboration between visualisation and simulation and data analysis.

Without using the "visual analytics" title explicitly, many VPH projects have combined visualisation with data analysis to achieve good results – see Figure 6.2 as an example of VPH2 [9]. The current interest of visual analytics, which has already been identified as a suitable technology within the Future & Emerging Technologies Programme, brings a useful opportunity to VPH, which has a high intrinsic demand for visual analytics.

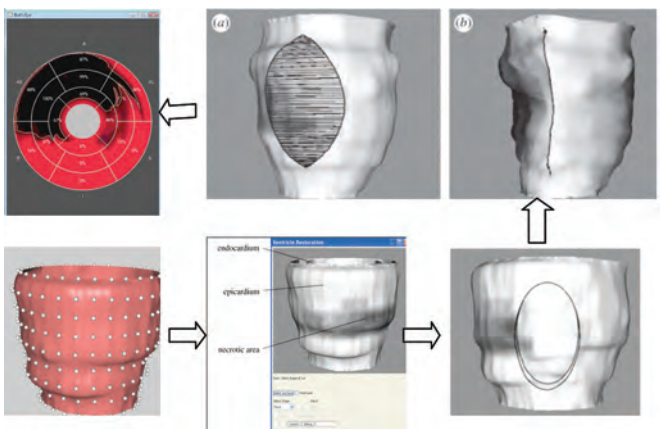


Figure 6.2: VPH2: Simulation & Visualisation of Left Ventricle Surgery: Cutting, Stitching and Deformation.

Another example of visual data mining is to generate hypotheses on biological networks and to reveal hidden patterns from microarray data. A developed system, VisGene [10], is illustrated in Figure 6.3. It allows biologists to integrate their knowledge into data mining by perceiving the process of correlation search, grouping and network modelling. Through good visualisations, they can see how the initial set of correlations between the genes is generated, how the grouping is formed based on the correlations, and how the Dynamic Bayesian Networks is produced. Based on the observations, the biologists can integrate their knowledge into the data mining and correct any errors by paying more attention to uncertain areas, changing algorithm parameters or switching to new data mining methods. Existing data mining only works out limited areas compared to the large number of genes in the network. The visual exploration of gene series will be able to help biologists to further alleviate other parts of the genes in the network and gain insights into the exploration of further deep sequencing.

Figure 6.3: The VisGene System[10].

One of the key features that has led to the success of visual analytics is the move from traditional confirmatory data analysis to exploratory data analysis by making the data analysis process more transparent to users, thus making them more deeply involved in the data exploration. This is clear contrast to the "black-box" strategy used in many of the existing VPH projects, in which clinicians and domain scientists employ intelligent data analysis techniques without being exposed to much of the content inside the box. The black box strategy often comes with strong limitations. In fact, a significant level of adaptation, adjustment and even extension to the standard techniques is usually essential to achieve the desired outcome in a specific application.

Thus, opening the black box to make the data analysis transparent is of great importance and visualisation plays a key role in this, as diagrams and pictures are very effective in terms of helping human comprehension. In comparison to the traditional visualisation techniques adopted in post processing, the visualisation in visual analytics needs to place more emphasis on the portrayal of the data analysis process and the associated uncertainties, indicating the key elements in the data analysis pipeline and highlighting the main areas of interest (e.g. areas with a high uncertainty) that should draw the user's attention.

It is expected that the application of visual analytics will allow us to uncover unexpected results and hidden truths within the dataset, probably as a result of the user involvement.

Impact on Biomedicine

As mentioned above, research in visual analytics will provide effective tools to allow VPH to achieve their goals in terms of being descriptive, integrative and predictive, and subsequently produce a profound impact on biomedicine. While visual analytics has been applied in many application areas, its adoption in biomedicine is currently very limited, and it is clear that many benefits could accrue from a systematic and well researched introduction of visual analytic approaches in VPH.

General Impact

Visual analytics has already been very successful in a number of cases, such as in the simulation of climate models, homeland security systems, etc. However, in general, many application cases so far implemented have shown only loose integration between visualisation and scientific computing and have not taken full advantage of the power of visual analytics.

At the moment, visual analytics research is mainly application driven – more applications need to be developed before we are able to address the underlying fundamental questions of visual analytics, such as what is the best architecture model for visual analytics, what are the best data structures and representations, what is the best way of integration, etc. The high demands for visual analytics in VPH make it an excellent application area for visual analytics to explore, and the knowledge gained could have a significant impact on developing a rigorous model by which visual analytics can operate more generally.

References

1. D. Keim, J. Kohlhammer, G. Ellis, F. Mansmann, VisMaster: Mastering the Information Age – Solving Problems with Visual Analytics, Eurographics Association, 2010
2. J.J. Thomas, K.A. Cook, Illuminating the Path: The Research and Development Agenda for Visual Analytics, IEEE, 2005
3. http://www.livinghuman.org/
4. http://www.vph2.eu/
5. http://www.vphop.eu/
6. http://bodybrowser.googlelabs.com/
7. M. Viceconti, G.J. Clapworthy, D. Testi, F. Taddei and N.J.B. McFarlane, Multimodal Fusion of Biomedical Data at Different Temporal and Dimensional Scales, Computer Methods and Programs in Biomedicine, 102(3):227-237, 2011
8. J. Urquiza-Fuentes, J.Á. Velázquez-Iturbide, A Survey of Successful Evaluations of Program Visualization and Algorithm Animation System, ACM Trans. on Computing Education, 9(2), 2009
9. N.J.B. McFarlane, X. Lin, Y. Zhao, G.J. Clapworthy, F. Dong, A. Redaelli, O. Parodi, D. Testi, Visualization and simulated surgery of the left ventricle in the virtual pathological heart of the Virtual Physiological Human, Interface Focus, Royal Society, 1(3):374-383, 2011
10. L. Zhang, W. Sheng, X. Liu, 3D Visualization of Gene Clusters and Networks, in Proceedings of Visualization and Data Analysis, 316-326, 2005

Editorial Team

Chapter Editors: Stefan Klein, Bjoern Menze

7

Data exploration and analysis in large heterogeneous imaging databases

Exploring

The Vision

Imaging plays a crucial role in analyzing human physiology at different spatial and temporal scales. Biomedical imaging provides structural and functional, biological and medical information, from the smallest scale (e.g. live cell imaging) to the largest scale (e.g. whole body imaging at the population level). Commonly, the physiological properties of interest cannot be directly measured, and some related images are acquired instead, from which the underlying physiological parameters are estimated in subsequent image processing steps.

Medical imaging is a typical "big data" problem[52]. Enormous amounts of biomedical imaging data are currently acquired; major data sources are clinical trials, population studies, screenings and clinical routine.

In many cases, the databases contain not only imaging data, but also genetic data, information on drug use, chronic diseases, diagnostic reports, patient history, etc. Commonly, the data are analysed solely within the narrow context of the study for which they have been acquired; often, however, their value would increase if they were combined with data acquired in different contexts. Infrastructural projects, such as the ESFRI EuroBioImaging initiative, aim to unlock all available data for use worldwide, by organising standardised storage, curation and access mechanisms. The general availability of such large amounts of data will then introduce the question: "How do we turn millions of raw images and other metadata into useful information?" Here the VPH initiative provides an answer: "By jointly integrating information of different sources through formal models of the human physiology."

Unfortunately, there is currently a complete lack of tools and methodologies to support the integrated analysis of imaging and associated data acquired in such studies. A particularly challenging aspect of this problem lies in the heterogeneity of the imaging data. This is caused by a lack of image generation and collection reference standards and the constant improvement in medical imaging sensor technology, which hampers the straightforward pooling of data acquired over time at different institutions. Moreover, the information usually originates from a wide range of imaging modalities, and all need to be analysed jointly in order to extract the underlying physiological parameters of interest.

Similar data exploration and data assimilation problems have been solved in related sciences leading, for example, to tremendous success of quantitative modelling in the geo-sciences. As a consequence, we envision a focused initiative for developing principled methods addressing this task in the same manner. It will enable us to explore the rich information present in the enormous pool of imaging data that will be produced in the future, with (almost) the same effort as analysing a relatively small, locally acquired, well-defined and homogeneous dataset. Solving this task will be a key to progress in biomedicine and, indeed, in any other image-related science.

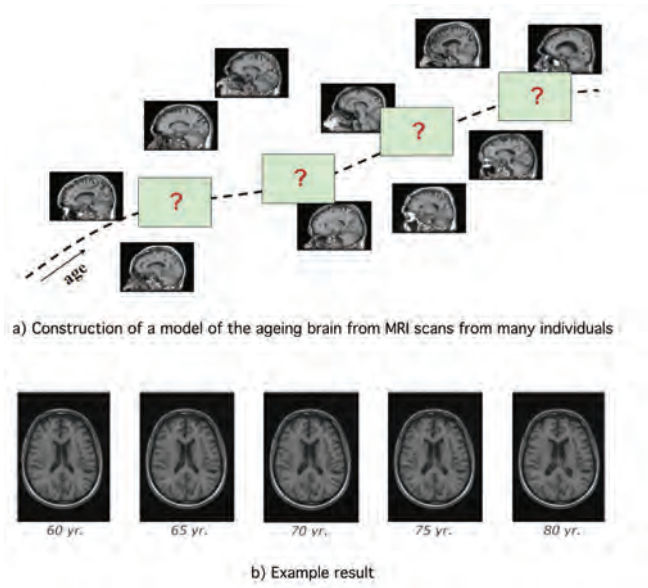


Figure 7.1: Anatomical brain change as a function of age. Construction of the model is presented in figure a). An empty frame indicates the population average at a certain age. Concatenation of all subsequent frames produces a model of the ageing brain. Applying the model to the scan of a (randomly selected) individual yields the result of figure b). Note the typical white matter atrophy that can be distinguished. This model is based on images from a single database (The Rotterdam Study, Erasmus MC, the Netherlands).

The challenge

Can we develop robust image processing and data analysis strategies that fully exploit the richness of information available in pooled large heterogeneous imaging databases, and that enable integrated modelling based on imaging data, genotype, patient history and other associated data?

Major advances will be needed in image processing, statistical inference, and pattern recognition. Below, we discuss the status quo and the required future technology for each of these topics.

Image processing

Manual analysis is no longer a viable option. Automated methods that allow for high-throughput analysis are required. Even the manual verification (by visual inspection) of automatically obtained results is becoming impractical. This means that the automated methods should be robust to changes in acquisition equipment and scanning protocols, and they should possess self-control capabilities.

Status quo

Despite an impressive amount of research on automated biomedical image analysis in the last decades, there is a lack of robust methods that can deal with the large variation of acquisition equipment and scanning protocols, without user interaction. Central to this problem are the following issues:

– Newly introduced algorithms are mostly accompanied by a few examples of successful application on a particular in-house dataset, but a systematic, large-scale, reproducible validation on publicly available reference data that reflects the variety of image characteristics encountered across institutions is usually missing.
– Many algorithms require the configuration of several user-defined parameters, which may have a large impact on performance criteria such as computation time, accuracy, and robustness. Tuning these parameters should be considered as "art" rather than science and should be performed by an expert with knowledge of the underlying algorithms. Principled approaches for tuning models given the data are not available, and testing model parameters for identifiability remains difficult.
– As imaging sensor technologies improve, the quality of the resulting images potentially improves, too. However, imaging sensors require tuning according to environment, lighting variations and patient cooperation. The quality of the resulting image affects the results of the processing algorithms. Institutions usually have their own in-house standards of image quality (usually through manual visual assessment); however, a standardised set of quality metrics for images of various modalities and dimensions is missing in most and is currently non-existent for applications across multi-site image databases.
– Basic physics teaches us that any measurement should be accompanied by an uncertainty estimate, i.e. a quantification of precision. In image processing, this aspect is usually ignored. In fact, this makes image processing more of a qualitative science than a truly quantitative one. As a consequence, tests on the validity of alternative modelling strategies – the successful basis for collaboration of experimental and theoretical research in physical sciences – cannot be performed easily in physiological modelling.

The biomedical image analysis community has become increasingly aware of the need for publicly available, annotated reference data and standardised evaluation measures, in order to address the first and the second points. Several large public databases have appeared, covering a wide range of acquisition equipment, scanning protocols, and anatomy; examples are the ADNI database for brain imaging (http://www.adni-info.org) and the Cardiac Atlas Project for heart imaging (http://www.cardiacatlas.org). Standardised evaluation frameworks consisting of reference data, evaluation criteria and evaluation software have, in recent years, been created in so-called "Grand Challenges" (http://www.grand-challenge.org) hosted at major international conferences, e.g. the "3D Segmentation in the Clinic" challenges at the Medical Image Computing and Computer Assisted Intervention (MICCAI) conferences. This has resulted in articles in top peer-reviewed journals in which multiple methods have been objectively compared. The framework of these grand challenges should also allow for more systematic investigation of the influence of user-defined parameters.

Almost no attention has been paid to the fourth issue: uncertainty estimation. Precision quantifies the impact of (both stochastic and systematic) measurement noise on the end result. This noise adds to – or even multiplies – modelling errors. To estimate precision and increase the quality of the physiological model, it is necessary to analyse the entire chain: the propagation of noise and artefacts from raw signal measurements, through reconstructed image intensities, to the result of post-processing steps should be understood. This would allow the user to further propagate the uncertainty to derived quantities. This last step is essential in the VPH context.

Required innovation

Automated image quality assessment. Before any processing, analysis or modelling can be successfully performed, the image must be assessed for its usability. A set of benchmarks is needed for the varied imaging modalities and dimensions (2D, 3D and 4D). Image quality assessment is also a contextual issue, as opposed to a "one size fits all" approach. An image that is good enough for global volume measurement may not be of sufficient quality for detailed shape modelling. The genetic and patient meta-data associated with the image must be accessible to help assess the usability of the image. Automating the process requires a full understanding and collaboration of the reference standards.

Providing uncertainty estimates. It might be acceptable that an algorithm is not always very accurate, as long as this is reflected in uncertainty bounds. For example, a human realises that in very noisy images, an exact edge location is hard to detect. By analysing how errors and noise propagate from the data to the final output, uncertainty estimates can be derived also for an algorithm. Uncertainty estimates provide a means for the necessary self-control in automated settings and high-throughput analysis.

Exploiting knowledge of the acquisition physics. By modelling the acquisition physics, imaging noise and artefacts can be taken into account, and pooling of data from different centres becomes feasible. An image processing algorithm should request the acquisition parameters (for example, echo and repetition time for MR images, whether

it was a 1.5T or 3T machine, high or low dose CT), and use these to derive the optimal settings for low-level algorithmic parameters whose effect is understood only by the developer of the algorithm. Knowledge of the data acquisition physics should also help to provide realistic uncertainty estimates and is a necessary prerequisite for automated image quality control.

Standardised validation. The impact of using different image analysis methods should be validated in terms of the final outcome (i.e., the model of physiological properties). Current validation is mostly focused on the immediate outcome of the processing algorithm (e.g., a contour that segments a particular organ in the human body, or a registration of two image modalities). Instead, independent measurements of physiological key variables are to be acquired and used for validation on selected benchmark data sets. Only by evaluating the methods within their entire physiological context, we can really say if a method is sufficiently accurate. This topic has a close relation to Chapter 4, "A software engineering approach to computational modelling research".

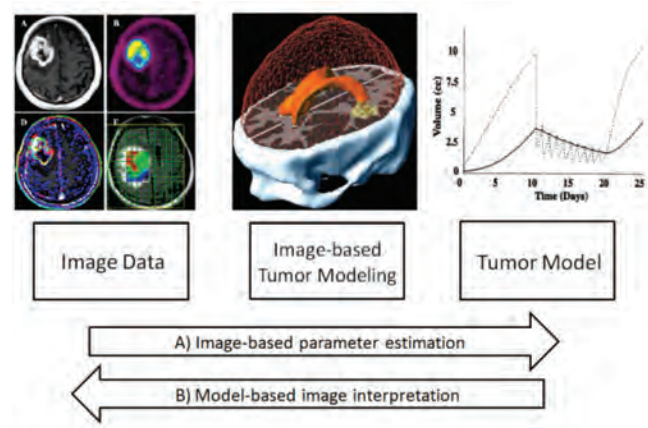


Figure 7.2: Image-based physiological modelling – an example of tumour growth. While sophisticated image processing routines and, for many applications, appropriate physiological models are available, tools for building principled interfaces between both fields of research are largely underdeveloped. (Figures by courtesy of MA Weber, Heidelberg University, and O Clatz, INRIA Sophia-Antipolis).

Model-based inference and image interpretation

Status quo

In many applications, there is a strict separation between image analysis and the use of relevant image information in physiological models. On the one hand, physiological models are developed without having specific modalities and imaging processes in mind. On the other, rather general image features are extracted to describe image content, without making specific use of a formal representation of the physiological process under study. This interface between image analysis and physiological modelling, however, is naturally crossed from both directions (Figure 7.2):

a) in the image-based adaptation of physiological models – for example in patient-specific clinical modelling, where descriptive model parameters are used in diagnostics;
b) in model-based image processing – for example in a task-specific image reconstruction in which physiological models provide guidance for a process-driven regularisation.

An example for the first direction is the segmentation of anatomical structures from an image when the exact, anatomically correct, geometric representation of these structures is a crucial input for subsequent biomechanical modelling. As example for the second direction, one may think of the vasculature of the brain, when one cannot decide from the image alone whether two vessels touch or are connected, due to insufficient resolution. In the first case, an image representation consistent with the subsequent model is highly relevant to model the spatiotemporal dynamics of the biomechanical process in a consistent manner. In the second case, a model may inform the image processing, using a model representation of the topology as prior knowledge of the vasculature, to arrive at the correct or the most probable representation (possibly highlighting several potential solutions), and even highlight disease-related deviations such as aneurysms.

a) From image to model parameter

Image processing serves as a filter: it automatically extracts features from images which may summarise the present information sufficiently well. This abstraction process, however, is driven by the visual interpretation of the image. Image descriptors (such as homogeneous regions, edge features, local descriptors of structure or texture, but also landmarks, statistical shape measures, or descriptive models of time courses in 4D images) are chosen as they somehow approximate image content at a lower dimensional representation, as they appear to be void of undesired variability or, simply, as they can be extracted through very efficient algorithms.

However, using these heuristic descriptors requires one to establish the relations to features used in the physiological modelling process. The same heuristic image features have to be derived artificially from the physiological model, and the interface between the physiological model and the image information will require extensive calibration and validation (in addition to the validation of the physiological model itself). Moreover, an extraction of image information only driven by image appearance may lose information, as the most obvious visual features do not always correlate well with the information required for inferring parameters of the physiological process. An image segmentation or a set

of characteristic landmarks, for example, is void of subtle information about image intensities; as another example, estimates without uncertainty quantification could appear to be precise even in situations in which they are wildly misleading.

Few image features are obtained through the inversion of physics-based models of the imaging process. Ideally, they would even have a strict interpretation as a physical property and would be used as direct input in physiological models – for example as a maps of water diffusion in tissue or local pressure. However, even here, there is a strict interface between estimating such parameter maps from imaging models und using them in a subsequent model of the physiological process under study. While it may be much easier to establish estimates of the uncertainty than for purely heuristic image descriptors, estimates are passed on without uncertainty estimates in most cases.

b) From model to sensitive image feature

In the other direction, models may explicitly be used to enhance features in images which are specific to the physiological process. This process may either be the (patho)physiological process under study, or it may be one of the covariates mentioned above, and the physiological model serves as a nuisance model allowing this undesired variation to be "regressed out". In the same way, the physiological model can be used as an integrative platform for fusing information from different imaging modalities (and even non-image information about physiology and function) in a very principled way. Models may help to decide about disambiguities when image information alone is not sufficient: As a concrete example, we can think of using a priori knowledge about the anatomical structures mapped by the image in order to extract an anatomically correct geometric representation of these structures, which is a crucial input for subsequent biomechanical modelling. In many cases, there is not enough information contained in the images alone to automatically decide about the correct representation. The spatial and temporal inter- and extra-polation of available information is one the greatest motivations for using functional models in image interpretation. The success of numerical simulation methods in related physical sciences – in particular, the geosciences – is to a large extent due to large efforts in data assimilation and parameter estimation.

When using a physiological model for similar purposes in image interpretation, however, it is not a small set of model parameters that is of particular interest, but the modelled state variable itself at image resolution. This poses a much harder problem, as model mismatch and uncertainties are not absorbed in model parameters (and their variation), but propagate to every observation in every pixel with very nonlocal effects. This direction – injecting physiological modelling into signal analysis and image interpretation – is not as well developed, partially due to a lack of appropriate physiological models, but also due to a lack of appropriate methods for estimating breakdown points under a given model hypothesis. Here again, uncertainty estimates of state and model parameters and their propagation to the imaging model would allow one to use available physiological models to a much larger extent, i.e., also in situations when they may only approximate the imaged process

Required innovation

Physiological models may span a wide range of growth processes, of disease progression, of processes in the healthy tissue, organs or subjects. In most applications, combining measurements from different imaging modalities is required, mapping complementary information about the underlying physiological process, informing biomechanical, biochemical or biophysical model components. Besides imaging data, other associated data – such as age, sex, diagnostic markers, genotypes – can also be integrated in a very natural fashion in the physiological model.

To make full use of this potential, it would be highly desirable to couple the inverse problem of the imaging process – often defined through acquisition physics – with the model of the underlying physiological process for solving for parameters and state variables of both models at the same time. In this, the principled treatment of uncertainty at every level, and the propagation of uncertainty through the entire signal processing and modelling chain would be of crucial importance both for the image-driven validation of the proposed models, but also for the model-driven analysis of image content. Inference in such a joint image and process model would provide a very general approach for propagating uncertainty from image observation to the physiological model. Reporting uncertainty at any level would also allow the identifiability of model parameters and model terms to be analysed, providing the means for a "hands on" use of a formal representation of complex physiological processes – and their problem-specific simplification – in image analysis and the processing of spatiotemporal signals.

From a modelling perspective, this should make inference more accurate, as more information is integrated into the model; and image information is closely related to features relevant to the model. From an image analysis perspective, formalised information on the expected spatiotemporal signal through physiological models may serve as a meaningful prior in signal processing, leading to a better interpretation of the acquired data and more informed decisions, for example, in biomedical diagnostics; it would make image analysis more robust and powerful as the physiological model provides a strong prior on the diagnostic task and serves as a platform for integrating all available image information.

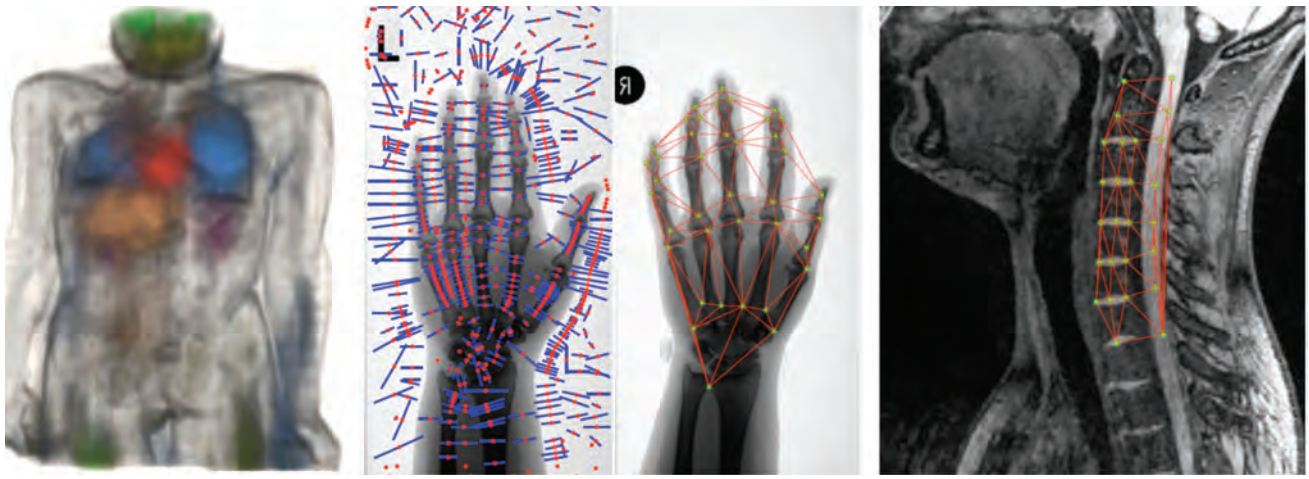


Figure 7.3: Pattern recognition and machine learning techniques in medical image analysis. Tools may be helpful for annotating large image data sets with little expert interaction, and for identifying structure from large scale databases using physiological prior knowledge (Images from http://groups.csail.mit.edu/vision/mcv2010; by courtesy of R. Donner, Medical University of Vienna, and A. Criminisi, Microsoft Research UK).

Pattern recognition and machine learning

Status quo

Supervised pattern recognition techniques that learn how to solve a certain image analysis task from a training set of examples are currently among the most powerful techniques in quantitative biomedical image analysis. Pattern recognition focuses on establishing statistical and empirical relations between well-described data. Relationships can be visualised to provide understanding of the properties of a given data set, or used to predict features of novel instances of the same process. They have been shown to be useful under specific diagnostic questions, e.g., in computer-aided-diagnostics, or in controlled experimental settings, such as the analysis of functional magnetic resonance images in neuroscience, or very reproducible experiments as in high throughput cell microscopy.

Unfortunately, the performance of supervised pattern recognition methods still largely depends on the choice of the training set, and faces significant problems when only small data sets are available to characterise high-dimensional image descriptors. Nevertheless, in the related field of computer vision, such data-driven pattern recognition and machine learning techniques – relying on large-scale "internet-wide" data sets, rather than explicit functional representations of image content – have become extremely successful, dominating nowadays over model-driven approaches in the automated interpretation of real-world images.

In the analysis of medical and biomedical image data sets, however, the variation from different imaging devices or technology, acquisition setting or natural variation of the human subjects, paired with highly specific analysis tasks still poses a much harder problem. In theory, an enormous amount of potential training data to develop pattern recognition based image analysis systems exists: similar images are analysed and similar decisions are made by many clinical experts every single day. Practical use of such methods, however, is currently limited by the fact that they require a sufficiently large training set of examples that are representative for a given task; as an example, it is often not possible to apply such algorithms directly to images taken with different imaging equipment. The need for a representative training set is therefore an enormous obstacle in practice which prohibits application both on databases where resources for establishing a ground truth are limited and in clinical practice where image analysis systems must work on a large variety of different image types and abnormalities.

Required innovation

One may address this general problem of data-driven learning techniques through three directions: A) by generating large annotated data sets representing selected key problems in physiological modelling sufficiently well; B) by exploiting unlabelled data more efficiently; or C) by designing smarter, model-based algorithms which are invariant to typical covariate and noise processes occurring in the observation of the physiological process under investigation. With each of these three directions, there is a common challenge, namely D) the requirement to scale algorithms to exceptionally large-scale databases. Special consideration should be taken of the natural variations of the human subjects that VPH aims to model, and the design of the learning approaches in the context of VPH should cope with, or be tolerant of, any sources of variation.

- **Large-scale annotation.** Learning from data requires, even to a larger extent than in model validation (see above), the availability of massive annotated data sets. This requires new tools for annotating such data, so that knowledge authoring becomes more efficient and standardised, enabling knowledge reuse. An example is active learning, in which a classifier is updated by iteratively proposing those samples to be labelled by an expert which increase the classification performance most, thus minimising the amount of user interaction. Also, using weak labels from other sources (for example context labels in medical reports), or developing approaches to the use of non-experts in annotating data (such as Amazon's "mechanical turk" which would possibly require a strict, automated quality control) may provide other directions for accessing large data sets in image analysis through pattern recognition techniques.
- **Exploiting unlabelled data.** Examples are semi-supervised learning and multiple-instance learning, which rely on few labelled examples and learn a good classifier by also using the available unlabeled data. This direction thus leads to algorithms that can more effectively generalise from a limited number of annotated samples, by exploiting the abundance of unlabeled data.
- **Model-based learning.** An interesting technique in the VPH context (or in any modelling science) is to consider structure in the feature space by combining statistical pattern recognition with formalised prior knowledge on a domain including knowledge on physiological structures or processes. Such model-based approaches may be used to guide the process of learning from data or to enable the use of unsupervised learning algorithms for modelling or recognition; either by formalising the process under study, or by formalising the variation for which robustness is pursued. This may also help in identifying functional relationships which describe, for example, the observed data by characteristic image intensity changes, shape variation in a population or in time, spatiotemporal trajectories on different spatial or temporal models. Ultimately, this would also provide a means for stepping beyond the simple modelling of statistical relations in pattern recognition to the explicit modelling of functional relations, and generate new hypotheses from image data acquired in the study of specific physiological processes.
- **Incremental or online learning.** Considering the magnitude of the VPH modelling objective, batch-learning methods would pose a significant overhead which, in many cases, would lead to considerable computational complexity. The need for incremental learning techniques that enable learning on an instance by instance basis is obvious; however current methods have limited scalability.

The VPH motivation

Modelling the physiological human requires measurements in the form of imaging data. Large 3D and 4D images are typically encountered, for which manual processing is not feasible. Automated analysis is necessary in such cases, which allows high-throughput processing. In most other chapters of this document, the need for standardised image processing algorithms is pointed out. In fact, the entire VPH vision relies on the availability of such analysis tools.

Imaging, and the subsequent image processing, is not a goal in itself. Imaging serves to provide insight into the morphology and function of tissue. Imaging is meant to quantify physiological properties that we seek to model. Commonly, the physiological properties of interest cannot be directly measured, and some related signal should be imaged, from which we can estimate the underlying physiological parameters in subsequent image processing steps. This requires knowledge about 1) the physics behind the imaging process and 2) the underlying physiological process (and a formal representation thereof).

To be able to accurately model the full complexity of the human body, a handful of datasets is not sufficient. The availability of large imaging databases is not just a luxury; it has become a necessity for serious advances in this field. However, without suitable technology to explore and analyse such databases, no gain should be expected.

Impact on Biomedicine

The availability of standards and methods for analysis of large heterogeneous imaging databases is essential not only for the VPH community, but also for progress in the field of biomedicine in general. At the basis of evidence-based medicine lies the principle that clinical decision making should be guided by scientific knowledge. Imaging data is increasingly forming the basis of such knowledge. Nowadays, there is a continuous quest for imaging biomarkers, which indicate the onset and progress of disease and measure the effect of treatment. Carefully validated imaging biomarkers can serve as surrogate outcome measures in clinical trials and treatment procedures, instead of conventional outcome measures that are more invasive to the patient, require much longer follow-up, or provide only qualitative information.

The VPH philosophy formalises the process to discover imaging biomarkers (knowledge) based on raw images and associated data. Image analysis and the subsequent inference of model parameters therein fulfil the crucial task of turning imaging data into quantitative measurements of physiological parameters.

General Impact

The need for data exploration and analysis in large heterogeneous imaging databases is not restricted to the biomedical domain. Images are everywhere. Astronomy, geology and machine vision are examples of domains in which imaging plays a key role. Robust automated image analysis is needed whenever visual inspection and manual assessment is no longer feasible. In all domains, we face the same questions: how do we extract relevant (physiological / astronomical / geological / product) information from the raw image data, with the final aim of turning this information into knowledge?

We are facing a shift of paradigm: the main challenge of research is no longer data acquisition, rather it is the abundance of the available data that needs to processed. A successful ICT approach to data exploration and analysis in such heterogeneous databases would be a key to major progress in any field in which models need to be extracted from large, heterogeneous, and pooled databases of measurements.

Editorial Team

Chapter Editors: Alfons Hoekstra, John Domingue

8

Web of models

Models

The Vision

One of the reasons that the VPH initiative advocates the use of predictive, computational models (translations of physiology into differential equation and other representation formalisms) is because we believe this is a good way to present a theory. Models provide a formal description of the theory; therefore, posing a biological hypothesis as a predictive model should provide an unambiguous and testable description of a proposed physiological or pathological pathway or mechanism. Because data and knowledge employed in models are distributed among geographically dispersed research groups and are represented in different formalisms, we need a framework with the necessary tools to combine and make interoperable these formalisations of our understanding of parts of biology and to reuse available models and experimental data. We call this framework a *web of models*.

A web of models will allow the creation of one or more *integrative models* and the reproduction of the complex workings of human biology, at different levels of detail. Simulated populations will be created. Medical problems *developed* by virtual populations will be studied according to models, relations among models, and individual risk profiles. Standard clinical trials will be complemented by virtual clinical trials. The web of models will support the provision of insights into the anatomy, pathology, disease progression and the most suitable treatments; and, through the development of clinical decision support systems, will allow the prediction of conditions and aid clinicians in decision making.

Within the scope of the VPH, there exist many different kinds of model and procedure. There is thus considerable difficulty in envisioning a system that could provide a general integration solution. However, progress has been made in specific areas, notably through the use of mark-up languages for model representation (e.g., SBML and CellML) and sharing (e.g., Linked Open Data). It will be important to provide flexibility for expansion, while beginning with more tractable sub-problems based on these technologies. Further development of mark-up languages is therefore required so that they can be used to represent and integrate all the desired kinds of model. The *COmputational Modeling in BIology NEtwork* will play a key role in overseeing the development of a set of open, interoperable and non-overlapping standards covering all the aspects of modelling in biology.

From a data sharing perspective, Linked Open Data is a technology for sharing data at web-scale and is based upon four simple principles outlined by Sir Tim Berners-Lee[53]:

- Use URIs as names for things.
- Use HTTP URIs so that people and machines can look up those names.
- When someone looks up a URI, provide useful information, using Web standards (RDF* and SPARQL).
- Include links to other URIs so that people and machines can discover more things.

As such, Linked Open Data can be seen as the simplest form of the Semantic Web with an emphasis on supporting the integration of data at the web-scale rather than on complex reasoning. The take-up of Linked Open Data has been significantly supported by a number of prominent players which we outline below.

The UK Government now has just over 7,400 datasets up at the government portal at data.gov.uk covering many areas of the UK government including the Cabinet Office, Ordnance Survey, healthcare, traffic data, crime data, etc. A screenshot of the data portal can be seen in Figure 8.1.

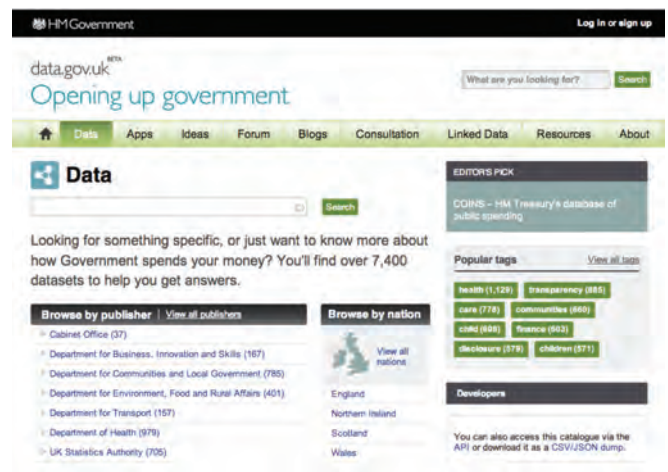


Figure 8.1: A screen snapshot of the UK Government portal at data.gov.uk showing the 7,400 datasets available.

[53] http://www.w3.org/DesignIssues/LinkedData.html

Within the portal, the wider community are encouraged to create Apps in a fashion similar to the Apple App Store based on the available data and currently around 150 of these exist see Figure 8.2. One of the benefits of the approach as outlined by the UK Government is that the effort of creating rich web pages to view governmental data has been delegated to the citizen and other motivated stakeholders.

In the domain of media, the BBC has also begun using Linked Data in a number of ways. For example, BBC Backstage[54] was a five-year initiative to open up the BBC by creating Linked Data for BBC programmes, enabling this to be browsed by programs. The BBC website for the 2010 FIFA World Cup was driven by Linked Data technologies[55],[56] which, according to the senior technical architect, facilitated more "agile modelling" compared to standard database techniques.

Figure 8.3 shows a screen snapshot from the website showing data for the English striker Wayne Rooney. This page is created on-the-fly using SPARQL queries over a

Figure 8.3: A screen snapshot of the BBC 2010 World Cup website showing information for the English striker Wayne Rooney[57]. Note that this page is created on the fly by querying a central RDF store which uses a world cup ontology.

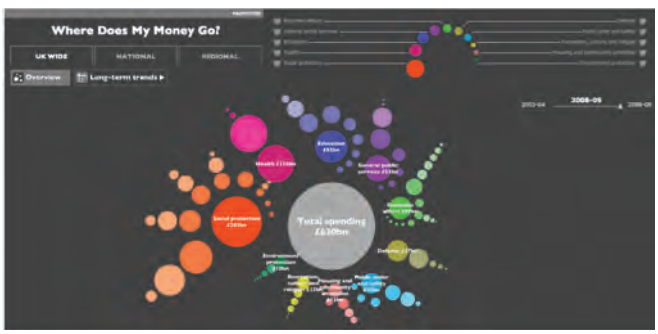


Figure 8.2: A screen snapshot of one of the Apps available at data.gov.uk. This App allows one to view how UK tax was spent by year (using the slider bar on the top right) and area (denoted by the colour and size of displayed circles).

Figure 8.4: A screen snapshot from the online news system of the Knowledge Media Institute at the Open University where Facebook users can say that they like a story.

[54] http://backstage.bbc.co.uk/
[55] http://www.bbc.co.uk/blogs/bbcinternet/2010/07/bbc_world_cup_2010_dynamic_sem.html
[56] http://www.bbc.co.uk/blogs/bbcinternet/2010/07/the_world_cup_and_a_call_to_ac.html
[57] This page can be found at http://news.bbc.co.uk/sport/football/world_cup_2010/groups_and_teams/team/england/wayne_rooney
[58] http://developers.facebook.com/docs/opengraph/

Figure 8.5: A screen snapshot the user's Facebook News Feed where the actions in the News website have been reflected.

central RDF store and this technology supported over a million hits per day at the peak of the tournament.

In the last few months a number of commercial companies have built sites around this feature. Levi's have a dedicated store which incorporates a "like" button for every product[59]. Also, Amazon have integrated their recommendation system to use Facebook profiles through Open Graph. Facebook have also recently integrated Open Graph into the Facebook SDK for the iPhone and Android platforms. When making the announcement, Facebook's CEO Mark Zuckerberg claimed that Open Graph is: "the most transformative thing we've ever done for the web"[60].

In late 2010, Google officially endorsed the use of the GoodRelations ontology to mark up online products and services. The general idea is that the markup is used to generate 'snippets' – the small segments of text shown as part of a search result. Figure 8.6 shows an example of this for the results of a search for a SCSI Controller card. One can see that compared to the first snippet (which does not use any markup) the second one contains the price of the item, review information, information on stock availability and a product photograph. GoodRelations also allows one to show the opening times of stores in the local vicinity which sell the desired item.

All of the above have contributed to the emerging Web of Data which is growing at a rate that matches the early part of the Web. Figure 8.7 shows a map of the data web as of September 2010.

Figure 8.6: Two screen snapshots showing how GoodRelations can improve how search results are displayed – top without and bottom with the ontology mark-up in use[61].

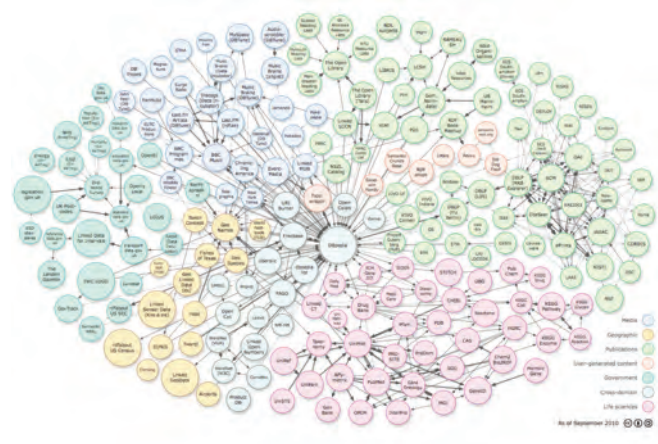


Figure 8.7: A map of the main datasets in the Web of Data as of September 2010. Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch http://lod-cloud.net.

This diagram shows over 25 billion RDF statements (specifically, subject-relation-object triples) with over 400 million links between them. In addition, the areas mentioned above the data sets also cover geography, publication data and life sciences.

Recent work has begun to consider how one can process the data above. Figure 8.8 (see next page) shows the main ways in which semantic service descriptions can be integrated with the linked data cloud. Firstly (8.8a), semantic annotation tools can be used to create RDF-based service descriptions that simply sit in the cloud, as for RDF descriptions of data following the Linked Data principles. Links from service descriptions to other linked data sets would be based primarily on descriptions of the types of data which form the inputs and output of a service. As the number of service descriptions grow, we would expect that the services

[59] http://store.levi.com/
[60] http://technology.timesonline.co.uk/tol/news/tech_and_web/the_web/article7104354.ece
[61] Taken from http://wiki.goodrelations-vocabulary.org/GoodRelations_for_Semantic_SEO

(a) (b) (c)

Figure 8.8: Showing three ways in which service descriptions can be integrated with the Linked Data Cloud.

descriptions themselves become linked expressing relations including: equivalent-functionality, super-functionality and sub-functionality.

Secondly (8.8b), we can create service descriptions for applications which directly use RDF data. We see now that communities are forming around linked data sets with the common goal of creating useful applications on top of the data. One of the best examples is based around the UK Open Government Data initiative mentioned earlier. As these are primarily web applications created over web resources using REST services, semantic service descriptions can help here to make services discoverable. As services may represent either application components or full applications, semantics can support both application developers and application end-users. Finally (8.8c), we can describe services that lift non-semantic services to the semantic level – by this we mean services that transform the output of services into RDF. The majority of semantic service frameworks include mechanisms for lifting (from syntax to semantics) and lowering (from semantics to syntax). Here we see a potential for discovering and reusing lowering and lifting mechanisms.

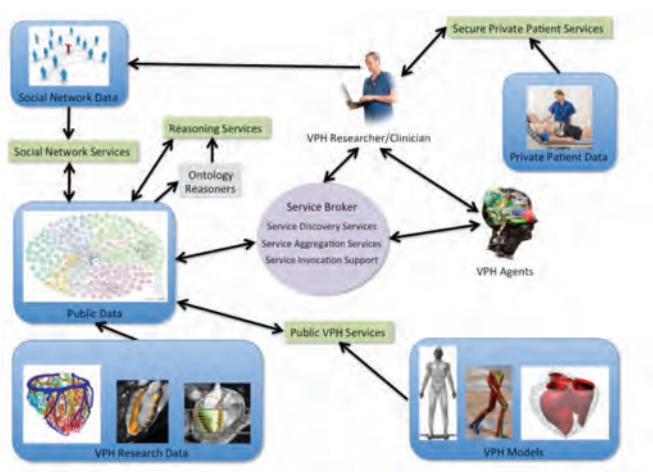


Figure 8.9: The VPH Web of Models vision.

The VPH Web of Models Vision

Building on top of the Linked Data substrate described above, our VPH Web of Models vision is captured in Figure 8.9. The elements describe the following.

- VPH Research Data is exposed as part of the Web of Data cloud following the four principles outlined above. In short, VPH research data should:
  a) when applicable, use the Web of Data – standard vocabularies such as geonames for places[62], FOAF for people[63];
  b) publish data sources within the Linked Open Data cloud; and
  c) provide SPARQL endpoints enabling queries to be conducted over the published data.

  It is important to note that the more links that are provided between data sets, the more useful the published data becomes.
- **Social Network Data** would be generated by the professional VPH community and contain information related to roles, organisational structures and current tasks. This data could be used to support: finding relevant experts on a specific VPH topic, communication and collaboration in joint ventures, the sharing of information related to the quality, applicability, authorship and ownership of VPH data and models.
- **Private Patient Data** would be available only through secure mechanisms to authorised users and organisations.
- **Ontology Reasoners,** based on Web standards such as OWL and RDF, would support complex and scalable reasoning over public VPH data. Note that the OWL-2

contains 3 profiles offering a range of different reasoning styles with a variety of performance characteristics. Specific reasoners would be required, depending on the reasoning task.

- **Social Network and Reasoning Services** would make the underlying data and functionality available to users, software programs, the Service Broker and VPH Agents. Descriptions of these services would reside within the Public Data space.
- **Secure Private Patient Services** in which authentication would be used to ensure that only privileged users gain access to the underlying data.
- **Public VPH Services** would make the VPH Research Data (as part of the general public data) and the VPH Models accessible to interested parties. As above, the descriptions of these services would reside within the Public Data space.
- **Service Broker** would support the management and use of all the VPH services using the service descriptions in the public data space. In particular, the broker would support the discovery of relevant services, the combination of services to support complex tasks and service invocation.
- **VPH Agents** would perform high level tasks using the service broker, for example, patient diagnosis on behalf of VPH researchers and clinicians. The main difference between the agent and the broker is that the VPH Agent would incorporate a high level planning and reasoning mechanism and thus be able to carry out complex tasks.
- **VPH Researcher/Clinician** would access the appropriate services either via the Service Broker or through the VPH Agent, and thus be shielded from most of the complexities involved in using and managing the underlying services. One exception to this would be secure services related to patient data.

The challenge

Following from the above we see a number of challenges for the VPH community:

- **Entering the Web of Data.** This first step is required to support all other activities. In addition to the technical issues involved there will exist additional hurdles related to data ownership, curation, quality, licensing. The Web of Data is currently mostly available under a Creative Commons Licence[64].
- **Supporting VPH Linked Services.** These will provide support for the creation of services that consume and produce VPH Linked Data; they would also be described as Linked Data.
- **Privacy, Provenance, Trust and Security.** These are issues especially if disparate data sources or services are combined from multiple owners or when combining public and private data.
- **Usability.** We have to ensure that the infrastructure created is usable by general clinicians or VPH researchers. This is will be non-trivial due to the nature of the technologies involved.

The VPH motivation

The integrative models at the basis of the VPH vision can be seen as a world wide web of reusable and combinable simulation-models, wrapped into agents or services. The concepts of *autonomous agents* and *Web services* will evolve in the future, but there will likely be equivalent concepts to which this vision can be applied. To implement this vision, the current semantic-Web paradigm is clearly not sufficient.

Part of knowledge can be captured only through models (of various kinds) that establish a dynamic relationship between portions of (the representation of) reality. Therefore, to make the Web of knowledge come true, a semantic Web is necessary, but not sufficient. We need a web of models, a way to capture dynamic knowledge into computational artefacts, which can then be composed into more-complex systems and used to process available data and generate new ones. This vision is much more generic than VPH, and can have a significant impact on ICT far beyond VPH. It requires communities with mature models, datasets and ontologies, and VPH is such a community, providing a strong background for the materialisation of the web-of-models vision.

If a web of models is in place, validated, predictive VPH-models can be exposed as semantic Web services, in which all inputs and outputs are well-defined concepts, relying on ontologies. For each model, sets of data stored somewhere are provided as input (either manually or automatically); ideally, all suitable sets of data that a semantic search-engine can find over the Internet are provided as input. Also, an input ontology could provide ways to restrict such a search. For example, if the input variable is "partial oxygen pressure in inhaled air", one can restrict it only to data taken from patients over 50, affected by COPD. Input variables of one model are connected to output variables of another model, thus weaving a web of models, with relations formally specified (using well defined ontologies). All the processing related to the execution of the models will initially be mainly local, but will shift towards remote, specialised locations over time.

Models in the *web* of models will undoubtedly have redundancy in the sense that there will be more than one model to represent a certain aspect of physiology. Some

[62] http://www.geonames.org/ontology/documentation.html
[63] http://www.foaf-project.org/
[64] http://creativecommons.org/
[65] www.synergy-copd.eu

selection criteria (with associated confidence/reputation levels) will need to be put in place to point to the best models, together with a matching mechanism to connect the models' input and output.

From the perspective of an individual, the web of models can be a place to interrogate personal data and monitor health evolution. Patients can visualise the details of their diseases and obtain expert information at various levels. Taking into account privacy concerns, data can also be shared with clinicians and other patients, and can be included in clinical studies and collaborative research studies. Data formats will be adapted in such a way that they can be used as input for different models, to predict conditions or to help clinicians making decisions.

For VPH to be realised, several requirements need to be satisfied. *Extensible/upgradable* ontologies for data and models are needed – and extensible/upgradable models are needed to prevent them becoming outdated with respect to relevant biological knowledge. Models need generic wrappers to be exposed as semantic Web services. A traceability mechanism is needed, to explain, at any point in time, how a certain dataset was generated. Data crawlers will discover new datasets, which will be used to match semantic searches and as input to existing models. Powerful user interfaces are needed for such a complex environment. A live data-network will be defined, created and maintained, with information stored remotely. Information repositories need to be updated and synchronised, and data inconsistencies need to be resolved. The level of automation with which data are analysed and inferences propagated need to be defined. Access to inferred knowledge need to be carefully regulated. Based on new personal data and on new and better integrated models, people will receive predictions and recommendations for a suitable time range.

Impact on Biomedicine

An interconnected web of models that are easily accessible, extendable and reliable can have a large impact in biomedicine in both the research and clinical setting.

Biomedical research

With the growing volume of data and models developed within the biomedical research community, it is becoming very hard for individual groups to re-use and track the developments of others. This is further complicated by the fact that different groups represent models in different ways and with different tools. The web-of-models vision will provide a framework and a methodology in which researchers can focus on problems specific to their domain and share their output (data or model) to other groups specialising in another related aspect.

Facilitating the integration of biological models will provide a deeper understanding of complex biological systems in which many properties are emergent (arising only when various parts of a system are combined). Efficiency will also be increased as the effects of new insights and developments in related research can be updated instantly without the need to re-implement a model in the researcher's own environment. A possible scenario is as follows. *Team A is working on a model of oxygen transport and delivery between the lungs and the tissues, and this is linked to a model developed by team B that describes lung heterogeneity via high-resolution CT images. An integration of these models would include inferred knowledge of how heterogeneity in diseased lungs affects the delivery of oxygen to tissues. If team B improves its model and publishes the updated model in the web of models, all users of that model will be notified that a new version exists and will be asked whether they want to update the connection of their models to the newer lung-heterogeneity model. This is a real example from an ongoing VPH project, Synergy-COPD[65], that is exploring the best way of integrating multi-scale models.*

Clinical Setting

The existence of a web of models can bring the objective of linking systems biology to systems medicine one step closer. Once an interconnected web of models (or part of it) has been verified and tested against new data and observations, it would then become established and trusted, and thus could be certified for clinical use. With a web of models, clinicians can tap into complex physiology, genomics and metabolomics models that can be used to run simulations on the lab results of real patients. The following are potential applications in the clinical setting:

– **Clinical decision support systems (CDSS) for treatment.** Incorporating complex biological models into decision-support tools will have a tremendous impact on medicine. These tools will provide essential assistance in the prediction and treatment of diseases by being able to simulate disease progression and different treatment scenarios for a particular patient's characteristics (e.g., MRI image, blood tests, DNA microarray analysis).

– **Early detection of diseases.** Access, via suitable interfaces, by the clinicians to biomedical models, which operate at the genomic and proteomic level, will help to predict the patient's risk and susceptibility to rare diseases, systemic diseases (such as COPD) or strongly mutagenic diseases (such as cancers).

– **Enhancement of medicine.** The web of models will be continuously updated and re-checked against newly published medical data; contradictions will be discovered;

and models can be improved to incorporate the new evidence. Patients will be able to contribute to public research by allowing the data in their EMR to be accessed anonymously by the web of models.

– **From research to practice.** The pace at which new findings in biomedical research arrive in the clinicians' office will be tremendously accelerated. Once a model used by a CDSS is updated and certified for clinical use, the system will be automatically notified of the new version and be offered the option of upgrading.

General Impact

The web of models vision will provide us with the ability to represent knowledge in an easily applicable, combinable and reusable way. The methodology is not just restricted to the biomedical domain but will have a large impact in virtually all sectors of information technology that deal with knowledge-intensive processes. R&D, financial and economic modelling, next-generation telecommunication services, public policies, travel and entertainment services are additional sectors in which the web of models would be very beneficial.

Conclusion

The big picture

In these eight chapters the panel of experts that participated in the VPH-FET consultation and consensus process delineated what are perceived by this research community as the grand challenges for blue-sky technological research associated with the full realisation of the VPH vision: that of a technology enabling a fully personalised and integrative investigation of the human pathophysiology by means of computer modelling.

While, if we read these detailed descriptions, the arguments seem very different from each other, looking at them from a distance, we can see a pattern emerging, a big picture that links these topics into a coherent whole.

Everything that is man made craves for simplicity. Occam's razor epitomises how the human mind reacts when faced with an infinitely complex reality: reduce it to simpler terms. The more we understand about life, the more we realise how far biological systems deviate from the ideal of functional perfection. In fact, processes in a biological system frequently involve a number of unnecessary or inefficient steps or imply a number of redundancies and they always embody a certain level of randomness. In other words, all biological processes are inherently highly complex.

For years, biomedical research has tried hard to reduce these processes to simpler terms, achieving important results in the process. However, it is becoming increasingly evident that such an approach can take us only so far in our quest for knowledge: there are biological processes that cannot be fully explained if we reduce them to their parts; in these, the whole affects the parts while, in turn, the parts affect the whole, and only by *embracing* the complexity can we achieve a full understanding of the systemic nature of these processes.

Ultimately, this is the essence of the VPH vision: develop a new framework of methods and technologies that make it possible for the human mind to embrace the complexity at the root of many important pathophysiological processes. Given our limitations, to embrace complexity we need to frame it, build it as collective endeavour and then tame it to our purposes.

Frame, build and tame complexity: these are the three characteristic elements that somehow connect all eight topics discussed above in a larger picture:

- **Framing the complexity:** integrative modelling of human pathophysiology. We first need to create a global reference framework, which will allow us to organise data, information and knowledge that we accumulate about any biological system in a coherent and integrated way. Then, we need to represent, properly and generically, the transition from a healthy state to the diseased condition. Last, but not least, we need to find ways of translating what we observe and learn from observing other species into this global body of knowledge about the human pathophysiology.
- **Building the complexity:** distributed and collaborative modelling of human physiology. Human physiology is infinitely complex, and can be tackled only as a collective endeavour. By building a web of predictive pathophysiological models that can be reused and composed freely, we can provide structure to this collective endeavour. Similarly, we can use information technology to radically modify the process of knowledge falsification, also a collaborative process, by ensuring that studies involving computational models of pathophysiology can also be reproduced and tested as well as physical experiments.
- **Taming the complexity:** interactive fruition of multi-input multi-output integrative models. However, all of the above would be futile if we cannot employ the integrative knowledge that is progressively and collectively assembled when trying to solve the problems of humanity. That is, we must also develop truly enabling technologies, capable of collecting and integrating disparate data and information somewhat related to the health status of an individual, and transform them into knowledge on how this status will evolve in the future. We need technology to explore this maze of complex knowledge to enable us to identify the proper leverage points or the specific actions that can heal or at least reduce the symptoms and the discomfort of our patients.

These are the goals that we believe the VPH should set itself over the coming years. To this end, we provide overleaf the text of a tentative Call for Proposals within the Future and Emerging Technologies Programme. We suggest that such a Call represents the most efficient and effective way in which the outcomes suggested by the various sections of this roadmap can be realised.

Tentative calls text

Title: Future and Emerging Technologies for the Virtual Physiological Human.

Summary: The Virtual Physiological Human (VPH) is a framework of methods and technologies that, once established, will allow the integrative investigation of the human body, seen as a whole. While for some specific clinical targets disease-specific VPH technology is already being deployed in research hospitals, the grand vision of a comprehensive VPH will be possible only through a generation of radically new technologies capable of enabling a completely new approach, which is holistic and personalised, to biomedical research and clinical practice.

Objective: Proposals should lay the foundations for radically new, potentially disruptive technologies for framing, building, and taming the complexity that a fully integrative approach to the human pathophysiology involves.

Target outcome:

a) Framing the complexity: integrative modelling of human pathophysiology
   - Global space-time reference system for human biomedical data integration
   - Transforming biological and physiological data across species
   - Disease modelling and virtual physiology

b) Building the complexity: distributed and collaborative modelling of human pathophysiology
   - A World Wide Web of predictive models of human pathophysiology
   - Reproducible and testable computational modelling research

c) Taming the complexity: interactive fruition of multi-input multi-output integrative models
   - Physio-environmental sensing and live modelling
   - Visual analytics for exploratory analysis of complex results sets
   - Data exploration and analysis in large heterogeneous imaging databases

Expected impact:

- New mathematical and computational formalisms for the integrative representation and modelling of human pathophysiology.
- Radically innovative technologies for the collaborative development, reuse, and distributed execution of predictive models of human pathophysiology composed into integrative models that effectively simulate complex biological processes in all their systemic interactions and across space-time scales.
- Radically innovative paradigms and technologies for the interactive fruition of multi-input, multi-output integrative models by the citizen/patient, the medical professional or the biomedical researcher.

Background documents: The "VPH-FET Research Roadmap: Advanced Technologies for the Future of the Virtual Physiological Human" elaborated by the Support Action *VPH-FET: Future and Emerging Technologies for the Virtual Physiological Human, FP7-ICT-258087.*

Annex

Annex #1: list of contributors

Over 150 experts worldwide participated to different extent in the consensus process that produced this research roadmap; their names and their affiliations are listed here in alphabetic order.

Surname	Name	Institution	Country
Aerts	Jan	Katholieke Universiteit Leuven	Belgium
Ali	Saad	Johnson & Johnson	U.S.A.
Antoniotti	Marco	Università Milano Bicocca	Italy
Arbenz	Peter	ETH Zurich	Switzerland
Arthur	Thomas	Proteus Associates	United Kingdom
Bacquet	Joël	European Commission, DGINFSO	Belgium
Badii	Atta	University of Reading	United Kingdom
Balasubramaniam	Chara	St. George's University of London	United Kingdom
Bertino	Elisa	Purdue University	U.S.A.
Bilsen	Greet	Katholieke Universiteit Leuven	Belgium
Bizer	Christian	Free University Berlin	Germany
Bogle	David	University College London	United Kingdom
Bourgine	Paul	École Polytechnique, Paris	France
Brook	Bindi	University of Nottingham	United Kingdom
Bullock	Seth	University of Southampton	United Kingdom
Calvetti	Daniela	Case Western University	U.S.A.
Castiglione	Filippo	National Research Council of Italy	Italy
Chapman	Tara	Université Libre de Bruxelles	Belgium
Christie	Richard	University of Auckland	New Zealand
Clapworthy	Gordon	University of Bedfordshire	United Kingdom
Clayton	Richard	University of Sheffield	United Kingdom
Cleather	Daniel	Imperial College London	United Kingdom
Connolly	Rory	US Environmental Protection Agency	U.S.A.
Cooper	Jonathan	University of Oxford	United Kingdom
Coppola	Gianfilippo	Imperial College London	United Kingdom
Coveney	Peter	University College London	United Kingdom
Dally	Bill	NVIDIA VP for research	
De Bono	Bernard	European Bioinformatics Institute	United Kingdom
de Bruijne	Marleen	University of Copenhagen, DK & Erasmus MC, NL	Denmark/The Netherlands
De Vlugt	Erwin	Delft University of Technology	The Netherlands
Deparis	Simone	EPFL Lausanne	Switzerland
Diaz	Vanessa	University College London	United Kingdom
Domingue	John	The Open University	United Kingdom
Dong	Feng	University of Bedfordshire	United Kingdom
Drasdo	Dirk	INRIA	France
Dubini	Gabriele	Politecnico di Milano	Italy
Dubois	Didier	Université Paul Sabatier	France
Ecabert	Olivier	Philips Research	
Ferrero	Josè	Universidad Politécnica de Valencia	Spain

Filipovic	Nenad	University of Kragujevac	Serbia
Fletcher	Katherine	University of Oxford	United Kingdom
Forgo	Nikolaus	University of Hannover	Germany
Fox	John	University of Oxford	United Kingdom
Frangi	Alejandro	Universidad Pompeu Fabra	Spain
Friboulet	Denis	Creatis Lyon	France
Gaggioli	Andrea	Catholic University of Milan	Italy
Garny	Alan	University of Oxford	United Kingdom
Georgopoulos	Panos	Rutgers University	U.S.A.
Gonzalez	Miguel	ALMA IT Systems	Spain
Gonzalez Garcia	Gerardo	University of Bedfordshire	United Kingdom
Gorissen	Dirk	University of Southampton	United Kingdom
Grzanka	Antoni	FEIT Warsaw	Poland
Guanyabens	Joan	Generalitat de Catalunya – Departament de Salut	Spain
Guntram	Berti	CMM	Germany
Gut	Ivo	PBC Barcelona	Spain
Hege	Hans-Christian	Zuse Institut Berlin	Germany
Heylighen	Francis	Vrije Universiteit Brussel	Belgium
Himeno	Ryutaro	RIKEN'S institutes & centres	Japan
Hoekstra	Alfons	University of Amsterdam	The Netherlands
Hoffman-Apitius	Martin	SCAI Institute – Fraunhofer	Germany
Hofstrat	Hans	Philips Research	
Hose	Rod	University of Sheffield	United Kingdom
Hunter	Peter	University of Auckland	New Zealand
Hurmusiadis	Vassilios	Epicardio Ltd	United Kingdom
Iakovidis	Dimitris	Technological Educational Institute of Lamia	Greece
Iakovidis	Ilias	European Commission, DGINFSO	Belgium
Iluyemi	Adesina	Mobile Diagnostic Services	United Kingdom
Jaecques	Siegfried	Katholieke Universiteit Leuven	Belgium
Jezek	Filip	Charles University Prague	Czech Republic
Jomier	Julien	Kitware Europe	France
Jordàn	Blanca	ATOS Orìgin	Spain
Kikinis	Ron	Brigham and Women's Hospital	United Kingdom
Klein	Stefan	Erasmus MC, Rotterdam	The Netherlands
Kohl	Peter	University of Oxford	United Kingdom
Kompos	Costis	Vodera Ltd	United Kingdom
Knuutila	Tarja	University of Helsinki	Finland
Krams	Rob	Imperial College London	United Kingdom
Krause	Mathias	Karlsruhe Institut of Technology	Germany
Lacroix	Damian	Institut de Bioenginyeria de Catalunya	Spain
Lalinde	William	O2Link	U.S.A./Spain
Lawford	Pat	University of Sheffield	United Kingdom
Le Novere	Nicolas	European Bioinformatics Institute	United Kingdom

Lemke	Heinz	IFCARS	Germany
Liaquat	Saad	University of the West of England	United Kingdom
Liò	Pietro	University of Cambridge	United Kingdom
Lonsdale	Guy	NEC Research Labs Europe	Germany
Loog	Marco	Delft University of Technology	United Kingdom
Maleckar	Molly	Simula	Norway
Marcus	J.T.	VU University Medical Centre	The Netherlands
Massey	Elizabeth	University of Lincoln	United Kingdom
Mazzà	Claudia	University of Rome "Foro Italico"	Italy
McClatchey	Richard	University of the West of England	United Kingdom
McFarlane	Nigel	University of Bedfordshire	United Kingdom
Menze	Bjoern	ETH Zurich	Switzerland
Merelli	Emanuela	University of Camerino	Italy
Meskers	Carel G.M.	Leiden University Medical Center	The Netherlands
Metspalu	Andres	University of Tartu	Estonia
Mirams	Gary	University of Oxford	United Kingdom
Mitsopolou	Evangelia	St. George's University of London	United Kingdom
Mitton	David	IFSTTAR	France
Moreau	Yves	Katholieke Universiteit Leuven	Belgium
Morley-Fletcher	Edwin	Lynkeus Srl	Italy
Motto	Santo	University of Catania	Italy
Müller	Ralph	ETH Zurich	Switzerland
Nelson	Martin	University of Nottingham	United Kingdom
Nickerson	David	University of Auckland	New Zealand
Niederer	Steven	King's College London	United Kingdom
Niessen	Wiro	Erasmus MC & Delft University of Technology	The Netherlands
Omholt	Stig	Norwegian University of Life Sciences	Norway
Otter	Tim	Crowley Davis Research Inc.	U.S.A.
Panetsos	Fivos	Universidad Complutense Madrid	Spain
Pappalardo	Francesco	University of Catania	Italy
Phillips	Andrew	Imperial College London	United Kingdom
Poulton	Terry	St. George's University of London	United Kingdom
Raggett	Dave	W3C	
Randall Thomas	Stephen	CNRS and University Paris-Sud 11	France
Remuzzi	Andrea	Mario Negri Institute	Italy
Rodriguez	Blanca	University of Oxford	United Kingdom
Rueckert	Daniel	Imperial College London	United Kingdom
Sanchez	Benjamin	Ministerio de Innovación y Ciencia de España	Spain
Sanchis	Teresa	Institut de Bioenginyeria de Catalunya	Spain
Sansom	Clare	Birkbeck University of London	United Kingdom
Santos	Andres	Universidad Politécnica de Madrid	Spain
Schouten	Alfred	Delft University of Technology	The Netherlands
Schroeder	Will	Kitware Europe	France

Schuppert	Andreas	RWTH Aachen University	Germany
Smallwood	Rod	University of Sheffield	United Kingdom
Smith	Nicholas	King's College London	United Kingdom
Stamatakos	Georgios	National Technical University of Athens	Greece
Stroetmann	Karl	Empirica	Germany
Stroetmann	Veli	Empirica	Germany
Summers	Ronald	Loughborough University	United Kingdom
Szekely	Gabor	ETH Zurich	Switzerland
Tegner	Jesper	Karolinska Institute	Sweden
Testi	Debora	B3C	Italy
Thiel	Rainer	Empirica	Germany
Thomas	Rhys	University of Leeds	United Kingdom
Toga	Arthur	Laboratory of Neuro Imaging, UCLA	U.S.A.
Turner	Philipp	University of Southampton	United Kingdom
Vallespin	Barbara	Universitat Pompeu Fabra	Spain
Van Riel	Natal	Eindhoven University of Technology	The Netherlands
Van Sint Jan	Serge	Université Libre de Bruxelles	Belgium
Varma	Susheel	University of Sheffield	United Kingdom
Vezin	Philippe	IFSTTAR	France
Viceconti	Marco	Istituto Ortopedico Rizzoli	Italy
Von Ammon	Rainer	Center of Information Technology Transfer GmbH	Germany
Waters	Sarah	University of Oxford	United Kingdom
Whiteley	Jonathan	University of Oxford	United Kingdom
Woodham	Luke	St. George's University of London	United Kingdom
Zang	Kaiyu	Imperial College London	United Kingdom
Zasada	Stefan	University College London	United Kingdom

Design by: breez design, 07979 945231