An introduction to uncertainty in the development of computational models of biological processes

Liesbet Geris[1-3] and David Gomez-Cabrero[4-5]

1 Biomechanics Research Unit, University of Liège, Chemin des Chevreuils 1 B52/3, 4000 Liège, Belgium

2 Prometheus, Division of Skeletal Tissue Engineering Leuven, KU Leuven, Herestraat 49 - box 813, 3000 Leuven, Belgium

3 Biomechanics Section, KU Leuven, Celestijnenlaan 300 C – box 2419, 3001 Leuven, Belgium

4 Unit of Computational Medicine, Department of Medicine, Solna, Karolinska Institutet, Sweden.

5 Center for Molecular Medicine, Sweden

Abstract

This chapter aims to provide an introduction to the different ways in which uncertainty can be dealt with computational modelling of biological processes. The first step is model establishment under uncertainty. Once models have been established, data can further be used to select which of the proposed models best meets the predefined criteria. Subsequently, parameter values can be optimized for a specific model configuration. Sensitivity analyses allow to assess the influence of the previous choices on the model output. Additionally, model adaptation permits to focus on specific aspects of the model without losing its global predictive capacity. Finally, predictions with the established models should also consider the effect of uncertainty in the model development process.

1. Introduction

Computational modelling of biological processes is becoming a standard tool used in biomedical research groups. The amount of examples showing the added value of the computational modelling approach is increasing by the day [1-3], more so if we include all Systems Medicine approaches [4]. One of the biggest challenges in creating useful models is the way in which they deal with uncertainty – uncertainty related to the experimental data but also to the modelling choices.

Uncertainty is defined in the Oxford dictionary as '*the state of being uncertain*'. Uncertain in turn is defined as '*not able to be relied on; not known or definite*' . Wikipedia defines Uncertainty as '*a term used in subtly different ways in a number of fields, including philosophy, physics, statistics, economics, finance, insurance, psychology, sociology, engineering, and information science. It applies to predictions of future events, to physical measurements that are already made, or to the unknown. Uncertainty arises in partially observable and/or stochastic environments, as well as due to ignorance and/or indolence*'[5]. Uncertainty in computational biomedicine can come from the experimental observations but can also be connected to the model itself either intrinsically (noise due variation in identically-regulated quantities within a single cell) or extrinsically (noise due variation in identically-regulated quantities between different cell) [6]. Throughout this book various ways of dealing with uncertainty are discussed. The structure of this chapter (and the whole book) follows that of the model development life cycle, starting with model establishment over parameter optimisation and model adaptation and ending with model prediction.

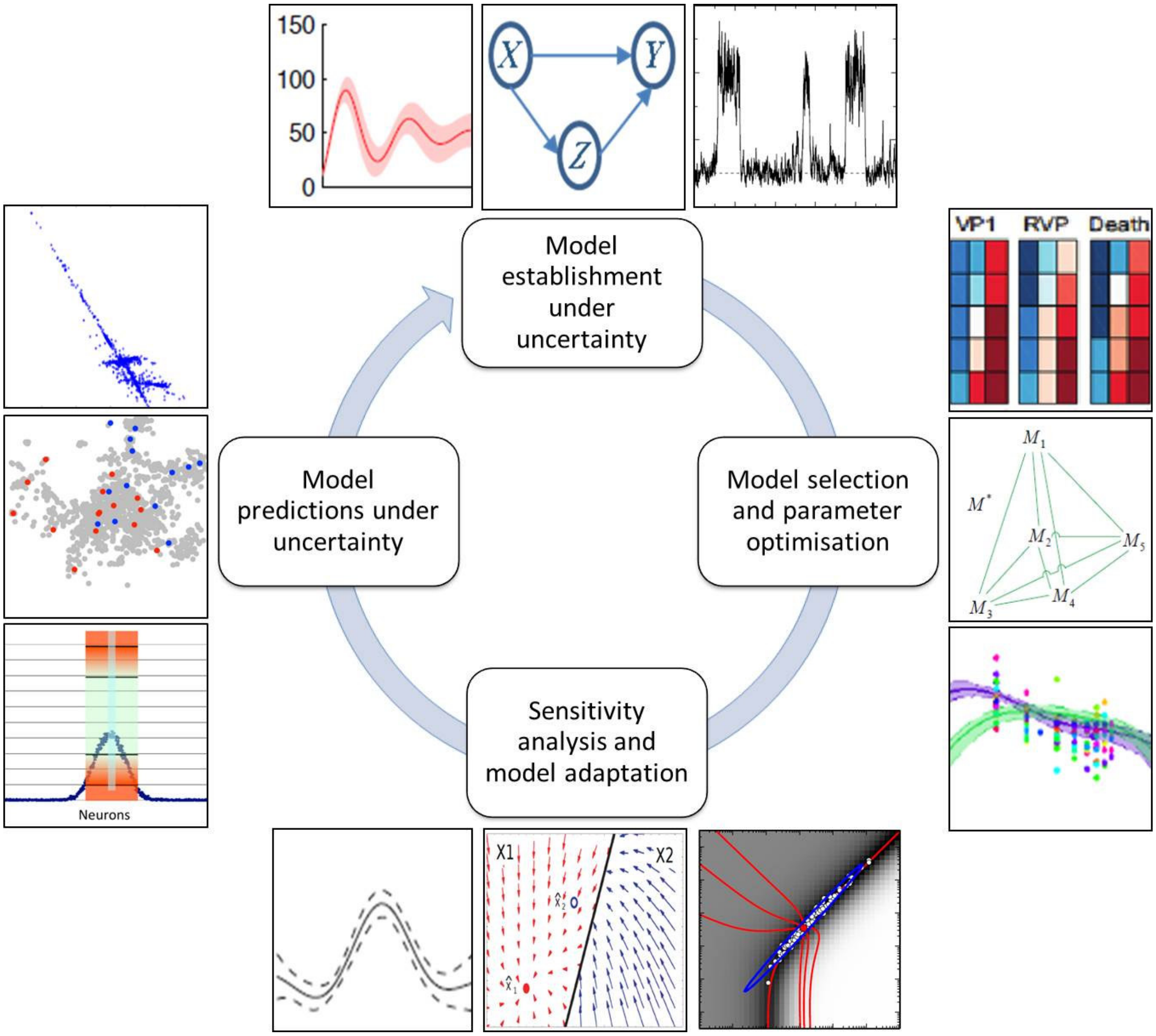
150
100
50
0
X
Y
Z
VP1
RVP
Death
Model establishment under uncertainty
Model selection and parameter optimisation
Sensitivity analysis and model adaptation
Model predictions under uncertainty
M_1
M^*
M_2
M_5
M_4
M_3
Neurons
X1
X2
$\hat{x}_2$
$\hat{x}_1$

Fig. 1: Schematic overview of the model development life cycle with a specific attention to the uncertainty in various forms: (i) establishment of a model under uncertainty, (ii) model selection and parameter optimization, (iii) sensitivity analysis and model adaptation, (iv) model predictions under uncertainty. Snapshot images have been taken from the following chapters in this volume (starting upper left corner going clockwise; C: chapter; F: figure): C2, F1 [6]; C3, F2 [8]; C4, F1 [9]; C6, F4 [11]; C9, F3 [7]; C10, F5 [14]; C11, F5 [15]; C12, F1 [17]; C13, F16 [16]; C16, F3 [20]; C16, F3 [20]; C17, F3 [21].

2. Model establishment under uncertainty

In order to *turn the ever increasingly available experimental (big) data into actionable knowledge*, mechanistic models are indispensable tools. They provide a conceptual and computational framework that allows for the interpretation and investigation of the experimentally observed behaviour and overcomes some of the limitations of classical statistical models in managing non-linear relations. Setting up these models however poses several challenges related to both the experimental and the modelling side.

Many experimental data sets tend to be noisy or incomplete and most often have not been collected with the specific intention of creating a model. Additionally, oversimplifications in model systems can obscure specific behaviours that have been observed experimentally. These uncertainties have an influence on the establishment of models. The chapter by Kirk et al. [6] provides an overview of the most prevalent problems in model establishment related to uncertainty in data and models, and proposes a number of strategies to tackle these problems. It furthermore discuss various techniques that have been put forward over the years to reverse engineer mechanistic models based on experimental data, each with their advantages and disadvantages. The chapter by Kirk et al. [6] additionally discusses some of the most common inverse techniques for this reverse engineering, including a more elaborate view on statistical inference techniques (additional discussion can be found in the chapter by Sunnåker and Stelling [7]). In the next chapter, Lagani et al [8] go into detail on a particular kind of statistical predictive models, namely the causal modelling. Computational Causal Discovery allows discovering causal relations with a limited set of interventions or manipulations. Causal modelling goes beyond traditional statistical predictive modelling as it provides the capacity to predict the effects of actions and interventions on a system, e.g., the effects of drug treatment, gene knock-out, or the induction of a mutation in the genome. This is in contrast to non-causal predictive modelling which is only valid when the system under study is observed under the same experimental conditions it was derived from and is not otherwise manipulated. Lagani et al [8] review the definition of causality and the basic concepts and principles of causal discovery, the nature of the underlying assumptions – particularly in relation to the uncertainty in the available data, potential pitfalls when applying the method, the most recent advances in the field and future directions. Both Kirk et al [6] and Lagani et al [7] are data-driven approaches to modelling under uncertainty.

A radically different way to establish models under uncertainty is the use of stochastic modelling and simulation techniques – which is particularly relevant when fluctuations become important. In Lejon & Samaey [9], the authors give a high-level overview of stochastic modelling techniques used for biological problems. They show the effect of stochasticity at different scales and different levels of description, and provide computationally reasonable solutions and algorithms for various problem types. They pay particular attention to the equivalence between the stochastic process that governs the evolution of individual agents and the deterministic behaviour of the related probability distributions.

3. Model selection and parameter optimisation

Once a number of potential models is established, a choice needs to be made on which model and which parameters are the most appropriate to use, given the experimental data that is available and the context in which the model will be used. This means that we need to understand the nature of the experimental data that is available to feed the models. After data acquisition, the use of data-driven modelling approaches allows to do a first processing of the data. Once a mechanistic model has been established, parameter estimation and optimisation can be performed in a variety of ways. Ultimately, when different modelling scenario's remain possible, specific tools can be used to determine which of these models is the most suitable, given the available data, the context and the preference of the modeller.

Western blotting, flow cytometry, protein mass spectrometry, DNA microarrays are just a few examples of a wide variety of experimental techniques frequently used in wet labs to gather data. In order to be useful for quantitative dynamic models, data needs to have a dimension of time as well as several perturbation experiments. The chapter by Bullinger-Schliemann- [10] provides an introduction to various experimental techniques that are frequently used in the model development life cycle, paying particular to the significance of single-cell versus population measurements. With the increase in availability of large and structured datasets, there has been a need to develop efficient data analysis techniques. Data-driven approaches, in contrast to mechanistic approaches, do not make assumptions on the underlying mechanisms. They are often used to process data to a more useful format and are particularly helpful in identifying biomarkers in large datasets. Shah et al [11] discuss a particular type of empirical models, namely the eigenvalue-based approaches. These approaches, including singular value decomposition, principle component analysis, and partial least squares regression, can identify important characteristics of big datasets through decomposition and dimensionality reduction. The chapter further discusses to way to deal with upscaling of these methods for understanding higher-order datasets (through tensor decomposition).

In the previous sections we described methodologies aimed to generate models from data; however data can also be gathered to define parameters in a model. When dealing with mechanistic models, the assignment of values to the parameter in a model, i.e. the parameter estimation, is a crucial step. Depending on the type and amount of data available, this can also be particularly time-consuming task. Over the last years, building on computational and algorithmic developments, many new tools have been developed to facilitate the parameter estimation step. A specific aspect of the estimation process is the optimisation, where the parameter space is determined that provides the most interesting results. An overview of these estimation and meta-heuristic optimisation techniques (simulated annealing, genetic algorithms, particle swarm optimisation and others) can be found in Samuelson et al [12]. As an alternative to these statistical-type parameter estimation methods, Tucker [13] describes parameter estimation via set inversion and constraint propagation techniques (interval methods). These techniques, based on set-valued computations combined with branch and bound steps, allow to examine entire sets of parameters and thus complete the global search within a finite number of steps. As the potential downside of interval methods is their relatively low speed, the author additionally shows how the method can be accelerated by set-valued constraint propagation, allowing for a considerable improvement of its efficiency. Samuelson et al [12] furthermore provide concepts and tools that allow the modellers to select the appropriate methodology for the specific scenario they are confronted with.

Once several models have been established, model assessment can help in identifying which model is the most appropriate for a given situation. Sunnåker and Stelling [7] discuss the most commonly used methods for model assessment of dynamical models, along with the underlying concepts and ideas. These methods include the information theoretic (e.g. the Akaike and deviance information criteria) and Bayesian approaches (e.g. posterior ratios for relative model probabilities from Bayes factors and the approximate Bayesian information criterion) as well as techniques such as cross-validation and bootstrapping. Bayesian model selection for biological dynamical systems is further elaborated by Hug et al [14], working with the Bayes factor computed by Thermodynamic Integration. Fundamentally different approaches to model selection (as compared to Bayesian approaches) are also treated, e.g. the minimum description length. All techniques are illustrated with examples ranging from simple, and sometimes analytically tractable problems, to medium sized models composed of ordinary differential equations. Information on how the most important results can be derived is provided in [7], alongside with a discussion on differences between methods ([7] and [14]) and how these methods can be employed in practice as there is no generally applicable method for model assessment that is valid in all situations.

4. Sensitivity analysis and model adaptation

Despite the techniques identified in the previous section, leading to the selection of the optimal model populated with the optimal parameter set, the uncertainty in the available data is often such that additional analyses of the parameter space and even model adaptation might necessary. Again, a variety of techniques is available to study the parameter space. Some techniques focus on the general character of the parameter space (e.g. sloppiness) for specific model types. Other techniques focus on specific pre-defined ranges in parameter space assessing the importance of their influence on the model results, i.e. sensitivity analyses. For over-parametrized or vary complex models, various simplification and reduction techniques have been developed to enable the understanding of the model's underlying core dynamics and a subsequent simplification of the model whilst maintaining its capability of capturing those core dynamics.

Exploring the parameter space can be a very challenging task due to its high dimension and complex structure. Mannakee et al [15] have shown that there exists a universal structure in the parameter space of models for nonlinear systems. More specifically, these models are often sloppy, with strong parameter correlations and an exponential range of parameter sensitivities all leading to good model behaviour. In their chapter [15], the authors review the evidence for universal sloppiness and its implications on parameter fitting and model prediction. They discuss how careful experimental design can lead to optimisation of parameter inference or general model behaviour (depending on the goals of the model and the modeller). They furthermore discuss the potential of transforming parameters to alleviate sloppiness. However, even when taking an information geometry perspective in order to have a parametrization-independent perspective on modelling, sloppiness arises and a deeper universal structure is revealed.

Rather than looking at the global parameter space, some methods specifically focus on a well-defined area, starting from specific intervals for all parameters present in the model (capturing the uncertainty of the parameters). 'Design of Experiments' (discussed by Van Schepdael et al [16]) is a technique originally developed to optimize physical experiments allowing to comprehensively determine the effect of parameters settings (individual parameters and their interactions) on the process with a minimal amount of experimental runs. For computational models, the limits on the specific parameter values that can be tested and the amount of runs that can be executed are generally less stringent but the amount of parameters (and especially their interactions) might be considerably higher than for physical experiments. The design of experiment approach for computational models allows choosing a minimum amount of parameter combinations that will result in a maximum amount of information about the computational model. The chapter by Van Schepdael et al [16] explains several designs and analysis methodologies.

The aforementioned methods all start from the fully developed model as derived in sections 2 and 3. However, with computational models of biological processes continuously increasing in size and model complexity (in part due to the data explosion in biology) it is increasingly difficult to obtain insights into what parts of a model generate a specific read-out. This hampers the correct interpretation of the model result and their use in e.g. the design of personalised therapies. The uncertainty in model structure and model parameters is a further complication. A solution to this dual problem of complexity and uncertainty is the systematic construction of simplified models from complex models. In their chapter, Eriksson et al [17] review different methods for simplification and reduction of models with particular focus on recent developments such as the iterative "tearing, zooming and simplifying" approach. This approach allows utilizing specific biological features such as modularity and robustness.

To wrap up the part on sensitivity analysis and model adaptation, two elaborate case studies are provided that investigate the sensitivity and effect of uncertainty on model outcome in the context of neural fields and bone mechanics. Laing et al [18] discuss the introduction of randomly chosen “frozen” spatial noise to their modelling system. The effect of inclusion of said noise on particular model outcomes, such as the occurrence of specific activity in a particular neural field model, is investigated and discussed. The second example is that of Mengoni et al. [19] who provide an overview of computational mechanical modelling of trabecular bone from a sensitivity analysis perspective. The effect of model development choices on model results is reviewed and analysed at different scales (from micro up to organ). As the focus is on models generated starting from Computed Tomography images, particular attention goes to the image processing effects, the mesh-related aspects and the computational representation of the boundary conditions.

5. Model predictions under uncertainty

With the model and its parameters all set, model predictions can be made that feedback to the experiments, closing the modelling life cycle. Model predictions will assist in advancing knowledge of the system under study in various ways. One such type of predictions is the identification of alternative explanations for and interpretations of the existing experimental data. Another type is the discovery of specific mechanisms in the simulation data. Both will lead to the formulation of additional experiments that need to be executed in order to validate (or falsify) the model’s observations.

Gomez-Cabrero et al [20] start from a very specific pre-frontal cortex working memory model and discuss issues related to non-uniqueness of parameter sets and the existence of various alternative solutions that can explain one particular experimental phenomenon. Using optimization techniques, they uncovered compensatory mechanisms in a subset of the parameters in the model, leading to the identification of hypothesis to be validated in dedicated experiments. On a more general note, Cedersund [21] provides an overview of various types of predictions that can be made – core predictions allowing to test the quality of the model or poorly determined predictions allowing to improve the overall well-determination of the model parameters. Even predictions that will not be tested experimentally can provide interesting insights into the studied model. In a medical context, reliability and accuracy of the predictions is important. Noteworthy is that this low degree of model uncertainty does not necessarily imply a similarly low degree of uncertainty on the model parameters. Such well-determined predictions are then also amenable to incorporation in larger supermodels (e.g. models of individual organs connected into a multi-organ model). Cedersund [21] subsequently provides an overview of the recent developments in the methods dealing with prediction uncertainty and discusses the price that needs to be paid when bothering with prediction uncertainty.

6. Conclusion

This chapter has provided a brief overview of the model development life cycle with a specific focus on uncertainty in the various stages: (i) establishment of a model under uncertainty, (ii) model selection and parameter optimization, (iii) sensitivity analysis and model adaptation, (iv) model predictions under uncertainty. Each of the following chapters in this book elaborates in a detailed way one or more facets of this development life cycle, with a specific attention to the incorporation of uncertainty in data and modelling. Taken together, the information provided in this book should allow modellers to start form experimental data, work through the different modelling life cycle steps and finally make predictions that can be verified experimentally. The last chapter, Gomez-Cabrero and Geris [22], provides also an overview of nowadays open challenges.

Acknowledgments

LG gratefully acknowledges funding the Belgian National Fund for Scientific Research (FNRS) grants FRFC 2.4564.12 and T.0254.13, as well as the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement n°279100.

References

1. Hunter P, Chapman T, Coveney PV, de Bono B, Diaz V, Fenner J, Frangi AF, Harris P, Hose R, Kohl P, Lawford P, McCormack K, Mendes M, Omholt S, Quarteroni A, Shublaq N, Skår J, Stroetmann K, Tegner J, Thomas SR, Tollis I, Tsamardinos I, van Beek JH, Viceconti M. (2013) *A vision and strategy for the virtual physiological human: 2012 update.* Interface Focus. 3(2):20130004. doi: 10.1098/rsfs.2013.0004.
2. Hunter P, Coveney PV, de Bono B, Diaz V, Fenner J, Frangi AF, Harris P, Hose R, Kohl P, Lawford P, McCormack K, Mendes M, Omholt S, Quarteroni A, Skår J, Tegner J, Randall Thomas S, Tollis I, Tsamardinos I, van Beek JH, Viceconti M. (2010) *A vision and strategy for the virtual physiological human in 2010 and beyond.* Philos Trans A Math Phys Eng Sci. 368(1920):2595-614. doi: 10.1098/rsta.2010.0048.
3. Clermont G1, Auffray C, Moreau Y, Rocke DM, Dalevi D, Dubhashi D, Marshall DR, Raasch P, Dehne F, Provero P, Tegner J, Aronow BJ, Langston MA, Benson M. (2009) *Bridging the gap between systems biology and medicine.* Genome Med. 1(9):88. doi: 10.1186/gm88.
4. Bousquet J1, Anto JM, Sterk PJ, Adcock IM, Chung KF, Roca J, Agusti A, Brightling C, Cambon-Thomsen A, Cesario A, Abdelhak S, Antonarakis SE, Avignon A, Ballabio A, Baraldi E, Baranov A, Bieber T, Bockaert J, Brahmachari S, Brambilla C, Bringer J, Dauzat M, Ernberg I, Fabbri L, Froguel P, Galas D, Gojobori T, Hunter P, Jorgensen C, Kauffmann F, Kourilsky P, Kowalski ML, Lancet D, Pen CL, Mallet J, Mayosi B, Mercier J, Metspalu A, Nadeau JH, Ninot G, Noble D, Oztürk M, Palkonen S, Préfaut C, Rabe K, Renard E, Roberts RG, Samolinski B, Schünemann HJ, Simon HU, Soares MB, Superti-Furga G, Tegner J, Verjovski-Almeida S, Wellstead P, Wolkenhauer O, Wouters E, Balling R, Brookes AJ, Charron D, Pison C, Chen Z, Hood L, Auffray C. (2011) *Systems medicine and integrated care to combat chronic noncommunicable diseases.* Genome Med. 3(7):43. doi: 10.1186/gm259.
5. http://en.wikipedia.org/wiki/Uncertainty. Consulted March 30th 2015.
6. P. Kirk, D. Silk, M. Stumpf (2015). *Reverse Engineering under uncertainty.* This Volume.
7. M. Sunnåker and J. Stelling (2015). *Model extension and model selection*. This Volume.
8. V. Lagani, G. Ball, J. Tegner, I. Tsamardinos (2015). *Probabilistic Computational Causal Discovery for Systems Biology*. This Volume
9. A. Lejon and G. Samaey (2015). *Macroscopic simulation of individual-based stochastic models for biological processes*. This Volume.
10. M. Bullinger-Schliemann, D. Fey, S. Livingstone, T. Bastogne and E. Bullinger (2015). *Parameter Estimation from the Experimental Side*. This Volume.
11. M. Shah, Z. Chitforoushzadeh and K. Janes (2015). *Statistical Data Analysis and Modeling*. This Volume.

12. O. Sameulsson, G. Cedersund, J. Tegner, D. Gomez-Cabrero (2015). *Parameter Fitting and the Optimization Problem*. This Volume.
13. W. Tucker (2015). *Interval Methods*. This Volume.
14. S. Hug, D. Schmidl, W. Bo Li, M.B. Greiter and F.J. Theis (2015). *Bayesian Model Selection methods and their application to biological ODE systems*. This Volume.
15. B.K. Mannakee, A.P. Ragsdale, M. Transtrum, R.N. Gutenkunst (2015). *Sloppiness and the geometry of parameter space*. This Volume.
16. A. Van Schepdael, A. Carlier & L. Geris (2015). *Sensitivity analysis by design of experiments*. This Volume.
17. O. Eriksson, O. Gutierrez Arenas and J. Tegnér (2015). *Simplification and reduction of large dynamical cellular models using biological constraints facilitate insights and derivation of model predictions.* This Volume.
18. C.R. Laing (2015) *Waves in spatially-disordered neural fields: a case study in uncertainty quantification*. This Volume.
19. M. Megnoni, S. Sikora, V. D'Otreppe, R.K. Wilcox and A.C. Jones (2015). *In-silico models of trabecular bone: a sensitivity analysis perspective*. This Volume.
20. D. Gomez-Cabrero, S. Ardid, M. Cano-Colino, J. Tegner, A Compte (2015) *Neuroswarm: a methodology to explore the constraints that function imposes on simulation parameters in large-scale networks of biological neurons.* This Volume
21. G. Cedersund (2015). *Prediction uncertainty estimation despite unidentifiability: an overview of recent developments.* This Volume.
22. D. Gomez-Cabrero, L. Geris (2015). *The Future of Modeling in Life Sciences*. This Volume.