models and methods for systems biology and systems medicine

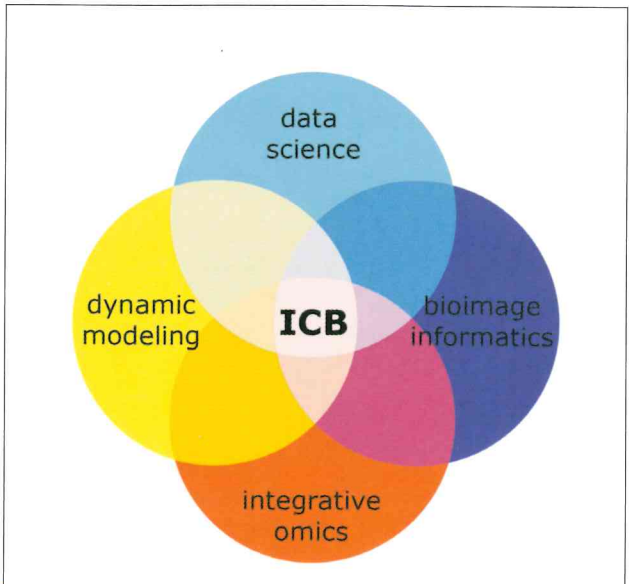
The Institute of Computational Biology at the Helmholtz Zentrum München

by Carsten Marr, Jan Hasenauer and Fabian J. Theis

The number of people with chronic illnesses continues to increase dramatically around the world. The key to understanding many such illnesses lies in the interaction of genetics, environmental factors and lifestyle. Innovations in biotechnology and the continual development of analytical methods permit us to obtain increasingly accurate measurements at the molecular, cellular and organismal level. This is associated with a rapid increase in the volume of data, which enables us to analyse a biological system from many different viewpoints. Today, for example, cells may be analysed using their genome, transcriptome, proteome or metabolome.

As a result, modern biological research requires mathematical and statistical methods to allow for efficient analyses of large amounts of data and the integration of various viewpoints. In addition, there is a growing need for statistical and mechanistic models to properly interpret the data obtained. In close collaboration with our experimental partners, our institute aims to establish analytical tools to enhance our understanding of diseases and their treatment options.

Research areas of the Institute of Computational Biology

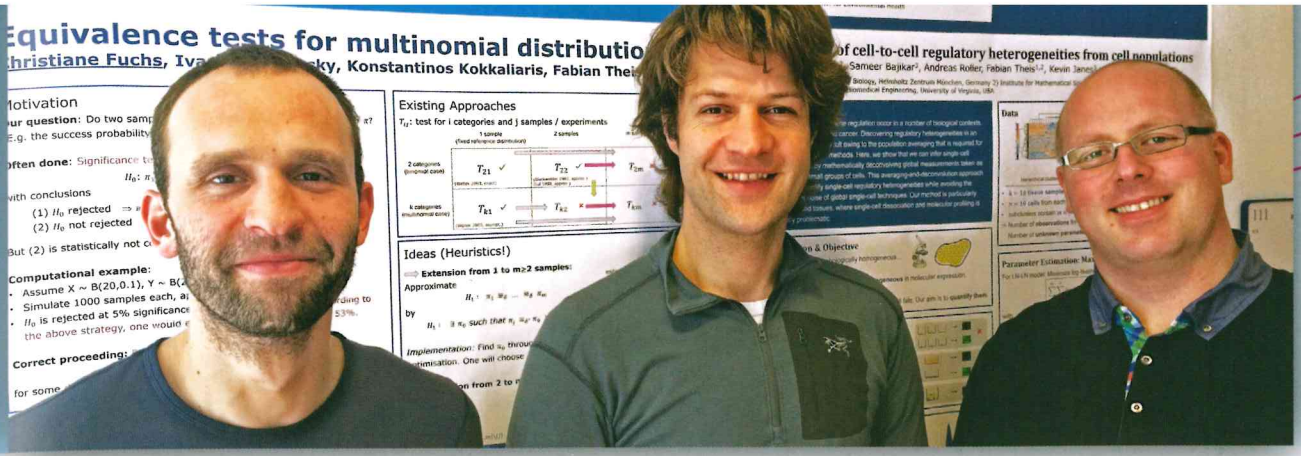


Graphic: ICB

The Institute of Computational Biology (ICB) resulted from the amalgamation of the Institute for Biometry and Biomathematics and the Research Group for Computational Modelling in Biology. The expertise of both groups was pooled in order to create new possibilities for the data-driven analyses of biological systems. Founded in 2013, the ICB is staffed by around 50 scientists and postgraduates. In addition to scientific work, our employees also lecture at Technische Universität München and supervise Master's and Bachelor's dissertations in the fields of mathematics, statistics, information systems and bioinformatics. The ICB works together with theoretical, experimental and clinical research groups at a national and international level. In addition, it is also part of several national industrial partnerships.

Science at the ICB

The ICB develops models and methods for analysing data in systems biology and systems medicine. We analyse information on a variety of scales – from time series of individual cells to Omics data from large patient cohorts. In our ten research groups, we are developing new methods for biostatistics, bioinformatics, image processing and mechanistic modelling, as well as integrative Omics analyses and data science. We apply these to the modelling of cellular decisions and the quantification of gene-environment interactions in disease pathologies. This article describes three of these research projects in greater detail.



From left to right: Carsten Marr, Fabian Theis and Jan Hasenauer (Photo: ICB).

Analysing cell-to-cell variability using statistical methods

Biological systems are highly adaptive and therefore very variable. Individual cells of the same type may react in very different ways to the same stimulus. Thanks to technological advances in imaging and the miniaturisation of reaction volumes with microfluidics, the description and analysis of this cell-to-cell variability is a new and exciting field of research. The ICB works to describe heterogeneity in the cellular context, e. g. gene expression variations in a mixture of differentiated and undifferentiated cells, using both statistical and mechanistic models.

Cellular heterogeneity is an essential factor in a range of projects in developmental and stem cell biology, but also in oncology. For example, we are working on acquiring a better understanding of the initial stages of murine embryonic development. After three divisions, a mouse embryo consists of eight cells, which start to differentiate into different types of cells. Experiments provided data on gene expression in individual cells after each cell division, which in turn provided us with expression analyses for different cell types. In order to detect differences between the cell types, we projected the 48-dimensional space of the gene expressions from single-cell qPCR onto a two-dimensional subspace. Each cell profiled thus corresponds to a point in the plane. With the aid of this projection, we were able to analyse which cells are very close together and which genes are responsible for transitions between cell types. Previously, it was only possible to differentiate cells after six divisions using standard projections. However, using the non-linear expansion developed and adapted by us, which also allows for group affiliations in the projection, it is possible to see that the cells can be categorised as one of two sub-groups already after four divisions (Buettner *et al.*, 2012). In practice, we ascertained that the resolution of transcriptomic data at the single-cell level resulted in new artefacts that were "averaged out" in the relevant data at the population level. For example, similar cells in different phases of the cell cycle could have significantly different levels of expression. In partnership with our colleagues at EBI, we recently recommended a method based on variance analysis in order to compensate for relevant confounders, such as the cell cycle (Buettner *et al.*, 2015). Thanks to the combination of single-cell analyses with statistical models, cells could be grouped into sub-populations that would otherwise have remained undiscovered.

From the cell to the patient

Interestingly, the methods developed for single-cell data can also be used for completely different types of data, such as individual measurements in large patient collectives. One such example comes from the field of diabetes research, in partnership with experts at Helmholtz Zentrum München.

Diabetes mellitus has been classified as an international threat and epidemic by the United Nations and is thus one of the biggest challenges faced by western industrialised nations. The mechanisms causing the disease are largely unknown. Until now, the best way of predicting the risk of type 1 diabetes was by examining family medical history and HLA genotypes. As part of a collaborative project, we were recently able to identify weighted gene combinations using statistical analyses that enable us to better predict the risk of type 1 diabetes (Winkler *et al.*, 2014). Our risk model with ten selected genetic positions enables improved risk prediction and therefore better screening of children in observational and intervention studies.

In addition to lists of known genetic risk markers, the institute also works with large Omics data sets. For example,

The academic staff of ICB at the retreat 2014 (Photo: ICB).

we recently created metabolomics networks that are able to depict the interactions between metabolic molecules specific to a type of tissue or organism. These networks were then expanded using genome-wide associations with genetic polymorphisms in order to create large, integrated metabolic maps showing metabolic and genetic correlations (Shin *et al.*, 2014). We then used these for a variety of purposes, such as analysing phenotype associations of the metabolome in order to simplify the biological interpretation of large results lists.

From measuring heterogeneities to understanding mechanisms

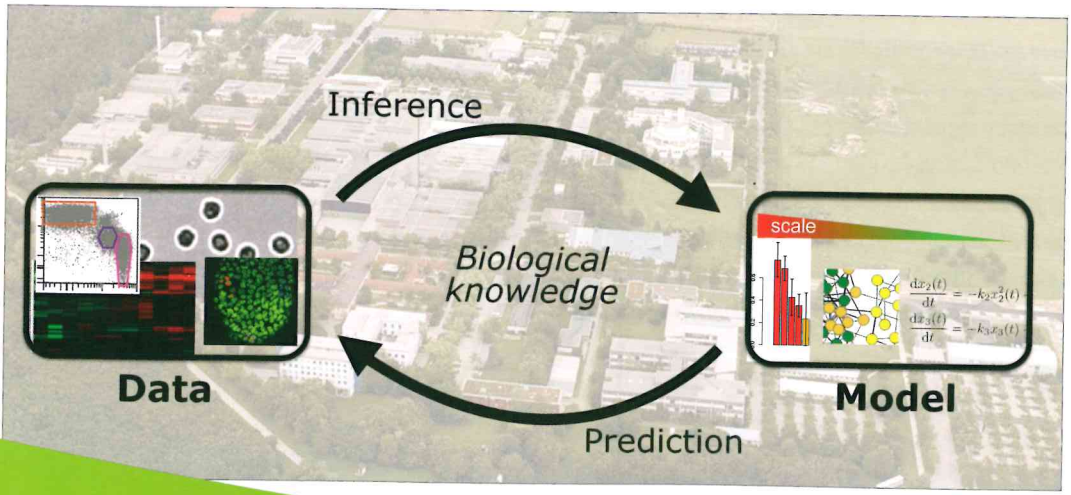
In order to better understand cause-and-effect chain, we use mechanistic dynamic models to analyse the *in vivo* characteristics of, for example, leukaemia, thereby promoting the mechanism-based stratification of carcinomas, or investigating cellular signal transduction. The development of deterministic and stochastic models is complemented here by tailored statistical evaluation methods. Together with other groups, we have developed algorithms that can be used to optimise models with several hundred parameters within hours. This allows the analysis of more complex data sets from a number of experiments.

We recently used such methods to identify various subgroups of neurones involved in transmitting and modulating pain (Hasenauer *et al.*, 2014). Through the combination of statistical and mechanistic models, we were able to determine the cause of differences between the subgroups, despite the fact that the cause had not been directly observed (Fig. 1). In similar projects, we worked with others to determine a potential target for the treatment of chronic pain, which is a major socio-economic issue.

Outlook

Innovative statistical methods and mechanistic modelling approaches are required in order to push ahead with establishing systems biology and systems medicine in the long term, both at our facilities and within Germany. Complex, high-dimensional, potentially longitudinal data sets are more and more available – partly within the specific project and partly via public databases – although clarification is still required on the questions of how to work with them and their integrative analysis in a wide range of projects. As a result, we want to develop tailored methods for complete analysis – from the cell to the patient – one step at a time, and push ahead with the development of multi-stage data-integration processes and genome-scale mechanistic models.

Data-based modelling at the Institute of Computational Biology



Graphic: ICB

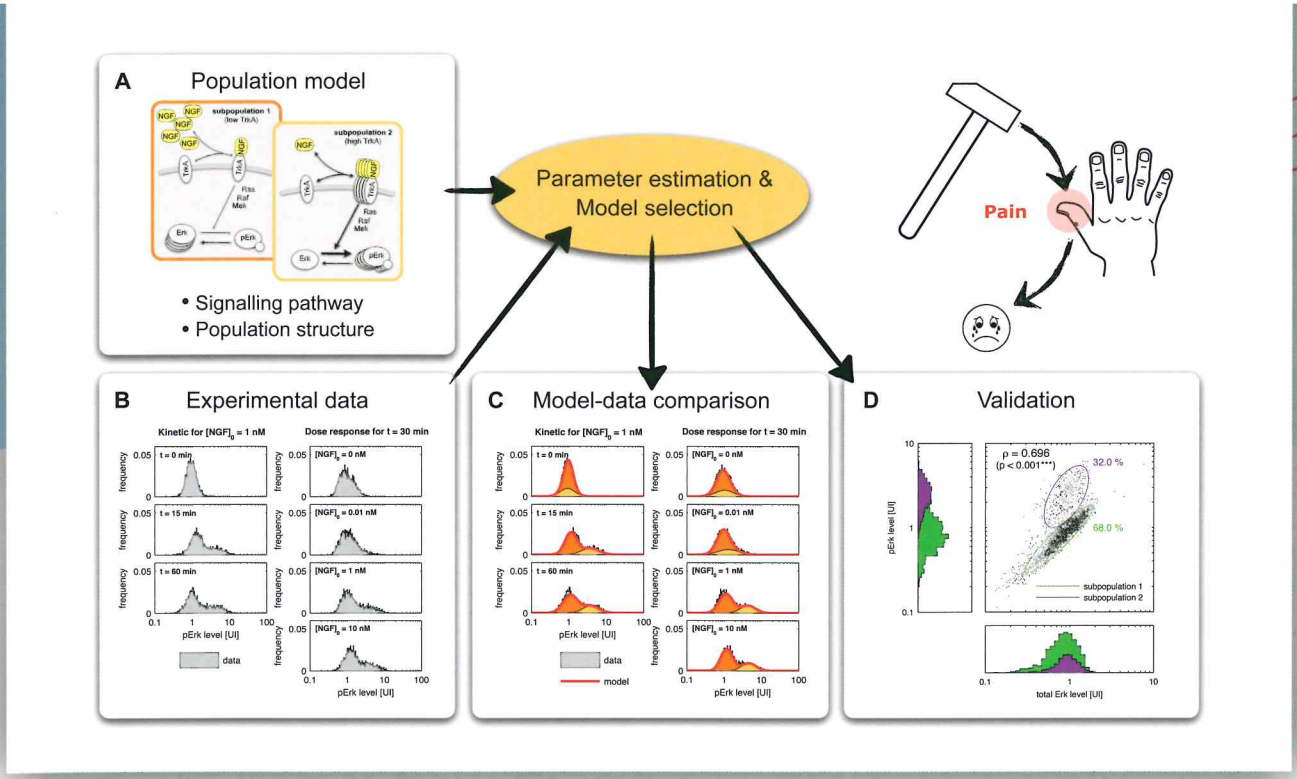


Figure 1:
Illustration of ODE-MM (Hasenauer *et al.*, 2014), a new modelling approach that draws on the advantages of synergies between mechanistic and statistical models. The intracellular dynamics of individual sub-populations can be described using mechanistic, ordinary differential equations. Cell-to-cell variability is depicted using mixture models. Using parameter estimation and model selection, these models (A) were adapted to experimental data (B), e. g. microscopy data. The resultant models (C) are reliable, with predictions of differences between cellular sub-populations, for example, having already been validated in a pain context (D).

References:

Buettner, F., Theis, F.J. (2012). A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst, Bioinformatics. 28 (2012) i626–i632.

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells, Nat Biotechnol 33(2):155-60.

Hasenauer, J, Hasenauer, C., Hucho, T., Theis, F.J. (2014). ODE constrained mixture modelling: A method for unraveling subpopulation structures and dynamics, PLoS Comput. Biol. 10, e1003686.

Shin, S.-Y., *et al.* (2014). An atlas of genetic influences on human blood metabolites, Nature Genetics, vol. 46, no. 6, pp. 543–550.

Winkler, C., Krumsiek, J., Buettner, F., Angermüller, C., Giannopoulou, E.Z., Theis, F.J., *et al.* (2014). Feature ranking of type 1 diabetes susceptibility genes improves prediction of type 1 diabetes, Diabetologia. 57, 2521–2529.

Contact:

Prof. Dr. Dr. Fabian Theis
fabian.theis@helmholtz-muenchen.de

Dr. Carsten Marr
carsten.marr@helmholtz-muenchen.de

Dr. Jan Hasenauer
jan.hasenauer@helmholtz-muenchen.de

Helmholtz Zentrum München – German Research Center for Environmental Health
Institute of Computational Biology
Neuherberg

Technische Universität München
Center for Mathematics
Chair of Mathematical Modeling of Biological Systems
Garching

www.helmholtz-muenchen.de/icb