Infrastructure for Systems Biology Europe

Work Package 4: Data Generation Centres

Deliverable No: 4.3

Strategic report on new infrastructure requirements

Dimitris Thanos-BRFAA

Hans Westerhoff-UNIMAN

Jutta Steinkoetter-MDC

Thomas Hoefer-DKFZ

Angela Oberthuer-UHEI

James Sharpe-CRG

Nicolas Le Novère -EMBL-EBI

Vitor Martins dos Santos-WUR

Jens Nielsen-Chalmers

Main/responsible Author(s): Dimitris Thanos, Sissy Kolyva, Ioannis Michalopoulos

Institution: Biomedical Research Foundation, Academy of Athens-Greece

Project funded by the European Commission under the Seventh Framework program for Research and Technological Development

Project ref. no.	INFRA-2012-2.2.4: 312455
Project title	ISBE – Infrastructure for Systems Biology Europe
Nature of Deliverable	R= Report
Contractual date of delivery	Month 30
Actual date of delivery	Month 36
Deliverable number	D4.3
Deliverable title	Strategic report on new infrastructure requirements
Dissemination Level	PU
Status & version	Version 3
Number of pages	16
WP relevant to deliverable	WP4
Lead Participant	BRFAA, Dimitris Thanos
Author(s)	S.Kolyva, M. Merika
Project coordinator	Richard Kitney
EC Project Officer	Keji-Alex Adunmo

Dissemination level: PU = Public, RE = Restricted to a group specified by the Consortium (including Commission services), PP = Restricted to other programme participants (including Commission Services), CO= Confidential, only for members of the Consortium (including the Commission Services)

Nature of Deliverable: P= Prototype, R= Report, D=Demonstrator, O = Other.

Contents

Overview 4

Data Generation Centres Action Plan 5

ISBE Data Generation Centres: Methods and Resources 8

Data Generation Centres building phase 11

Application-Focused Infrastructure 12

Overview

Most biological processes involve network interactions between multiple genes, proteins and environmental variables. The complexity of these interactions in time and space is enormous, creating highly individual and variable responses. Attempting to unravel and understand the dynamics of these processes requires the collection and integration of experimentally-derived, quantitative, systems-wide data on the state, dynamics and variability of living cells, organs, organisms and populations. Handling and interpreting these diverse data sets requires the use of a variety and complex set of computational, mathematical and statistical modelling techniques and can only be achieved with a critical mass in both the experimental and quantitative sciences. Achieving this effectively to deliver the ultimate goal of understanding how biological function_emerges from interacting biological components is *the* major challenge for modern biology and lies at the core of evolving systems approaches.

Systems Biology is at the current frontier of life sciences, representing a highly interdisciplinary approach for understanding biological complexity in health and disease. In the last decades it has become clear that all life sciences relate to systems. The focus of medicine today is the systematic and systemic treatment of all multifactorial diseases, as compared to traditional single molecule targeted approaches. Therefore, new Systems Biology Infrastructures capable of supporting technology transfer as well as of translating scientific and medical discoveries into medical applications and other applied concepts will improve both our basic understanding of life and Medicine and enhance the economic potential of Europe. ISBE is an integrated research infrastructure that will exploit existing synergies and will create new opportunities for efficient research coordination and collaboration in the field of Systems Biology. ISBE will make available cutting-edge technologies in experimental and computational systems analysis to the wider community of European life scientists. Centres specialising in specific experimental and/or computational technologies will combine in a variety of operational clusters to create the backbone of the infrastructure. These centres will contribute to ISBE not just through specific projects, but also through the training of researchers - and by acting as hubs for stimulating further technology development. Additionally, the ISBE centres will catalyse the broader integration of the quantitative sciences of physics, mathematics and engineering with biology by providing a unique environment where scientists from all these disciplines meet, work together and educate each other.

ISBE will offer services and resources that are useful for all branches of the life sciences, independent of the type of organism or biological system studied. Systems biology creates a strong unifying and theory-based foundation in the life sciences. The different types of services include resources for stewardship and standardisation to make data, models and tools re-usable, modelling of biological systems based on integration of diverse data sets and the facilitation of model-compliant data generation.

Regarding the quality of data suitable for systems biology work, it is noted that most existing biological data sets are unsuitable for systems biology modelling: they are incomplete, unannotated, or have been acquired for other purposes and not necessary under physiological conditions. Researchers active in the systems biology field generally require precise data obtained under defined experimental conditions.

To address this, ISBE will facilitate the generation of data suitable for systems biology through: (i) the development of community standards and best practices for maps, data, tools, models and SOPS; (ii) the provision of brokerage services to bring researchers in contact with external research infrastructures or institutes with experimental design and data generation capabilities in compliance with ISBE standards; and (iii) support in the experimental design phase and throughout the data generation, integration, modelling and model validation process, in order to obtain model-compliant data.

Data Generation Centres Action Plan

Building a European research infrastructure such as ISBE is a long-term process. The ISBE planning began in 2009 with the submission of a proposal to ESFRI. This resulted in ISBE's inclusion in the ESFRI Roadmap 2010. This enabled ISBE to obtain funding through the EU FP7 Infrastructure programme to fund its Preparatory Phase (2012-2015) in which 23 research institutions and funding organisations in 11 countries collaborated. Since systems biology is a relatively new and rapidly developing field in the life sciences, the Preparatory Phase was primarily intended to:

- map and analyse the European systems biology community
- explore the potential user and provider-base in academia, hospitals, clinics and industry, including SMEs
- develop a simple, easily accessible and affordable structure for operation of the research infrastructure

In the forthcoming Interim Phase (2015-2018) the ISBE research infrastructure will be built in a stepwise fashion until it reaches full functionality in 2018, that is, 10 years after the ISBE initiative was conceived. ISBE is envisaged as an infrastructure where clusters of research groups from institutions/centres will focus their various and specialised expertise on discovery-directed and hypothesis-driven research, and/or by contributing to the development of underpinning technologies. The combined expertise and facilities of the ISBE will serve the European Research Area by functioning as the entity for addressing important scientific problems, by disseminating technologies and by providing open and active access to data, software and experimental and modelling facilities (e.g. to the extent of enabling external researchers to perform experiments in a relevant cluster either directly or real-time-through-web). Although ISBE institutions/centres will have complementary activities, each will typically support the following: *de novo* data generation, data extraction from all pre-existing sources, data management and curation, data analysis, model extraction from literature, *de novo* model generation and validation, visualisation and modelling, dynamic interaction of models and data, model-driven experimental design, and training. As the available data in the public data-bases are of no real use for systems biology work (unsuitable), there is an urgent need for the production and the subsequent curation of new high throughput highly quantitative data that will be suitable for systems biology work. Importantly, the ISBE Data Generation Centres will take a leading role in exploring, developing and establishing the necessary standards for experimentation, data curation and modelling in systems biology, critical for ensuring the delivery of reliable and consistent data across the infrastructure. A key deliverable of ISBE as a whole, and of DGCs is to bring focus to these disparate efforts by identifying, structuring and supporting large-scale research projects on major areas of urgent need in medicine, biotechnology, the biosciences and the economy - for example, human physiology and ageing, complex diseases, bioenergy and bio-manufacturing. Such a focused European infrastructure will offer a single point of contact for access to a network of best laboratory practice in Systems Biology, unique in the world, stimulating contacts with non-EU consortia from academia, industry and regulatory agencies across the globe, as well as other related infrastructures and programmes within the EU.

ISBE aims to provide to both existing and new users of systems approaches. This community continues to grow. The ESFRI Assessment Group Report recognised the requirement for an Interim Phase to develop pan-European infrastructures, which allows the identification of founding centres, and provides time for negotiations to finalise a longer-term legal framework. While it is not possible to state exactly how large ISBE would ultimately become, it is expected that within the first couple of years of operation at least 10 national centres will start to deliver services under the ISBE umbrella. ISBE is in the process of identifying potential nSBCs (National Systems Biology Centres) and developing the process for establishing the interim boards and panels necessary to support ISBE beyond the end of the Preparatory Phase period, as of August 2015. ISBE will continue discussions with potential centres and associated funding bodies, including through ISBE's audit of the potential provider-base during the preparatory phase. This will be followed by an open nomination process in 2015 that will require candidate centres to detail existing or planned financial national support, together with endorsement by both the host institution and an associated national research ministry or funding body.

In order to develop plans to allow the start of operations, many, if not all, of the standing bodies and associated management and secretariat functions will be established. The Interim Phase will follow immediately after the end of the Preparatory Phase (August 2015) and is expected to last 2 to 3 years. Regarding the Data Generation Centres, the goal is to develop a comprehensive plan based on the survey performed on task 4.1 to identify which of the existing infrastructures are capable of providing access to high throughput technologies. In parallel, the gaps in the available infrastructure and the needs for building new centres will be defined. Furthermore, the level of commitment of the existing and new infrastructures to the future of ISBE will be defined.

The analysis of the data presented in Deliverable 4.2 indicated the existence of a large number of research facilities spread through out Northern Europe that could in principle provide Data Generation Services to the European Systems Biology Community. However, it remains unclear the degree of commitment in providing services in the context of ISBE, an issue that should be addressed in the Interim phase 2015-2017. An additional issue that should be addressed is the small number of facilities capable of producing high quality Systems Biology-grade data in the rest of Europe. ***Thus, it is important to evaluate the capacities and abilities of existing Data Generation infrastructures to service the community. Otherwise, there is a pressing need for upgrading the existing facilities and build new ones at selected sites.***

The conceptual challenges for such a plan of distributed Data Generation Centres will be based on:

*t*echnological issues: currently, technological deficits exist in mathematics, computation and experimentation. Experimentally, there is a demand for high-throughput and other technologies that will help populate quantitative models. Further improvements in measurements at all scales are needed to obtain biologically relevant quantitative data. It is particularly important to sample living systems dynamically, at multiple scales, if realistic models are to be constructed. The systems biology Centres are encouraged to develop innovative approaches to address these and other technological challenges.
DGCs will constantly examine and evaluate the existing state-of-the-art available technologies to determine whether future technological and scientific developments in the scientific areas of systems biology should integrate these novel technologies.

- There are also continuing challenges in applying mathematics to the complexity of living systems. Growing volumes of data from diverse high-throughput experiments provide unprecedented opportunities for computational biologists, but also challenges in data storage, analysis, archiving and visualization.

- *Training, Outreach and Organizational Challenges*. Building cohesive multi-disciplinary research teams by integrating expertise across traditional disciplinary boundaries is not a simple undertaking. There is a continuing need to disseminate knowledge widely through outreach activities to the research community through appropriate conferences, personnel exchanges, and sharing of resources. Importantly, the emergence of a new science demands an adequate, diverse workforce of appropriately trained scientists. The future leaders of systems biology research will be knowledgeable and skilled in both experiment and computation. Innovation in research training therefore, is a significant task of the Data Generation Centres.

ISBE Data Generation Centres: Methods and Resources

To develop the basis for a Pan-European systems biology infrastructure like ISBE, we must consider how centres can work together to establish the Infrastructure and stimulate the integration of the systems biology community in the areas of quantitative data production, training, data management and sharing, modelling, technology development, and academic-industrial links.

Databases play a major role in systems biology research, with a central role currently played by expression array and omics data and its analysis. Data resources need to be developed in a sustainable manner, and the continued funding of database infrastructures is a major problem. Analysing the stores data requires quantitative models, for which the data must be organized to aid in determining causal relationships. Modellers have the choice between going through lots of papers published over past several decades on individual experiments or using the available high throughput datasets. Datasets mostly have static rather than time-dependent data, but do reflect what can happen and they are important as a kind of scaffold information. Data and infrastructure requirements include e.g. Biochemical definitions of different molecules composition, factors that read, write and erase, Analysis of Post Translational Modifications: e.g. histone and non-histone proteins. Characterization of proteins (complexes) involved in transcription and chromatin organization Identification of enzymes that can be drug targets. Deciphering the epigenome of a cell, i.e. to determine where in the genome and when in normal or cancer cells/tissue a given factor binds or epigenetic mark is present etc.

Methods and resources for generating this data include:

Comprehensive and quantitative omics-level experimental approaches including Next Generation Sequencing, epigenomics, microarrays/deep transcript sequencing, polysome profiling, proteomics, RNAi screens etc. Progress in quantitative and real-time single cell approaches, microscopy and image analysis, flow/FACS etc. Agreed and implemented standards for sample and data collection and representation. Clinical proteomics for diagnosis and prognosis. A repository of high-quality and well-annotated clinical sample collections Improved high-precision omics and imaging measurements from molecules to cells.

Major initiatives are in progress to gather extremely wide ranges of genetic variation data for both somatic and germline variations that are disease relevant, e.g. the International Cancer Genome Consortium and the Cancer Genome Atlas (TCGA). The data from these initiatives should be used to provide scientific input to bioinformatics and systems biology analyses to understand the biology of cancer or its response to drugs. Relevant systems biology modelling requires the development of new technologies and computational/ mathematical tools driven by the biology requirements, with human diversity studies is one of the most obvious needs. At another level, the integration of clinical and medical data and resources (clinical records), epidemiological information with molecular information (genomics) is a very obvious need. Furthermore, a common information infrastructure for data exchange, analysis and modelling using standardized data integration and meta-analysis methodologies all of which are derived from concepts that support the optimal design of experiments. For example, if we choose cancer as a model system to be studied by systems biology-related approaches, it seems necessary to focus our attention on the identification of molecular differences between healthy and carcinoma cells. The problem is complex in view of the fact that molecules from many parallel signal transduction pathways are involved. Their functions seem to be controlled by multiple factors. Numerous nonlinear effects of regulatory feedbacks, pathway cross-talk and non-stationary biochemical processes complicate the understanding and prediction of these intracellular dynamics. Formal methods need to be developed that help identify subsystems (networks/pathways) which can be studied in focused experimental studies.

Mathematical methods should support the design of novel experiments that allow the distinction of alternative hypothesized network structures on the basis of experimental data. Related to these complications is the question whether so called inhibiting deregulated pathways affect the carcinoma cells so that the disease will go into remission. To answer such questions, we need to determine how important different enzymes are for signalling in a pathway, and for cell survival and growth. By comparing similar determinations for normal cells and cancer ones, we would be able to reveal

which enzymes or pathways make the most effective targets for carcinoma treatment. For these analyses, we especially need:
- Cell-context specific molecular interaction maps in cancer (cancer interactomes)
- Widely available experimental platforms for rapid biochemical validation
- A detailed ORFeome (protein and tagger protein expression)
- siRNA screening assays
- Instrumented cells (reporter genes for all genes in a cancer cell model)
- Unbiased hypotheses about oncogenic lesions and processes
- Cellular network based contexts for the integration of orthogonal data modalities, including gene expression, SNPs, gene copy number, epigenetic data, etc.
- Information on pathway synergy for therapeutic intervention
- Assembly and validation of cell-context specific molecular interaction maps for cancer cells (genes, proteins, miRNA, lipids, metabolites, etc.)

These analyses should:
- Provide a deeper understanding of causal relationships in cancer initiation, progression and treatment (data-driven functional and regulatory networks; cancer stem cells)
- Allow accurate prediction of disease and treatment outcomes (diagnosis and prognosis)
- Enable engineering of novel therapeutic interventions (molecular, cellular, physical)

At all stages of disease progression, major areas requiring modeling via systems and developmental biology methods include immune system reactions, angiogenesis and tumour progression. Key research areas include:
- Biomarkers: The use of organ-specific blood protein fingerprints for diagnosis
- Genomic data: Genomic data is central to making disease predictions (when integrated with environmental information)
- Emerging technologies for medicine: Next generation DNA sequencing, microfluidic /nanotechnology approaches to measuring proteins in complex mixtures; the creation of new chemistries for generating new protein-capture agents, single-cell analyses and new in vivo and in vitro imaging technologies
- A focus on one particular system, e.g. colon cancer or gliobastoma
- Comparisons of mouse models, cell lines and human samples

Modelling across scales is a major challenge. There are dynamic models needing a lot of data and higher order models which need different types of information. Experimentalists and modellers need to mutually discuss how to produce the right data and the right models and provide the right way to store the data. When people think of multi-scale modelling, the focus is on cells and organs.

From the above, it is quite obvious that systems Biology-based approaches usually require the acquirement of new data derived from experiments that have been

designed for this purpose. It is highly unlikely that the available data in the public databases will be useful for future systems biology research. Therefore, there is an urgent need for the establishment of ISBE-associated Data Generation Centres.

Data Generation Centres building phase

Road map to shape DGCs to a fully operational life sciences technology infrastructure hubs. These actions build on the strength of the founding parties and together with the stakeholders in the life sciences.

DGCs are developed by a consortium of stakeholders in the life sciences: Building a European life sciences technology agenda will only be successful if all stakeholders participate, including universities, funding agencies and industry. Important starting point is that DGCs build on extensive experience of life sciences technology communities in providing third parties access to expertise and infrastructure. Based on previous investments, active and well-structured communities in genomics, proteomics, metabolomics, imaging, bioinformatics, engineering etc. have been already developed in Europe.

Develop a DGCs technology agenda: A life sciences technology agenda based on the need for life sciences technologies in the different life sciences sectors in the next five years will be developed using the data of WP9 (Technology Watch). From this analysis, a realistic estimate should be made for the strategic investments both in equipment and in expertise that are required for the development of DGCs. The agenda will list the type and kind of infrastructure that could be shared by the red, green and white life sciences, as well as sector-specific components. It will take into account specific infrastructures that are essential for coping with the exploding data volumes in the life sciences, making links to the European ICT agenda. The life sciences technology agenda will set the scene for the development of DGCs.

First steps towards DGCs- matchmaker projects: As a first step towards the development of collaborative DGCs, at European and national level specific calls should set up for proposals to team up research projects that need high-end technology with centres (part of nSBCs) that are able and willing to provide state-of-the-art infrastructure facilities and the necessary expertise. If successful, project parties may arrange themselves a continuation, especially where pilot projects link to themes of larger EC programmes. This matchmaker approach will also provide DGCs with insights into issues related to the collaboration between DGCs in order to obtain data sets that can be integrated, as well as the data integration process itself. At the same time it should put DGCs on the map in the research community.

At the core of Systems Biology and especially in Systems Medicine is the close interaction between clinicians, basic researchers and theoretical partners. To exploit the ample knowledge available in the different disciplines and to jointly advance it to gain insights into disease promoting mechanisms, improve diagnosis and advance therapeutical options, mutual understanding and effective communication is essential. This requires a technological infrastructure that supports efficient integration of different levels of information and is tailored to the needs of the users. An important basis is the development of an agreement on data production, handling, storage and sharing and means to access the quality of knowledge utilized for the formulation of mathematical and computational models.

For basic research data, the awareness for the necessity of standardized procedures for data generation has been raised during the past years. To ensure comparability and the sustainability of data, it is essential that the procedures including assay performance should be standardized. Measurements are preformed at the cell population and single cell level as well as from the cellular up to the body scale. On the one hand, high throughput data for the genome, epigenome, proteome, transcriptome and metabolome is gathered and on the other hand very detailed quantitative and time as well as dose-resolved measurements are performed. Due to the variability of biological systems even apparently small alterations such as minor changes in temperature or pH can have a major impact on the behavior of a system. Usually multiple variations of protocols for assays have evolved and are favored in different laboratories. A key challenge is to distinguish major from minor differences and pragmatically agree on standard operating procedures that are widely accepted and applied. Furthermore, it is important to appropriately document the obtained data. Therefore, harmonization is required of annotations and units as well as of the metadata that is documented.

Application-Focused Infrastructure

The combined implications of integrating modern quantitative biological approaches with complex system modeling and computational infrastructures are unique to biology. Although some individual pieces of the integrated vision discussed above have been implemented and in some cases have been optimized for different platforms, the next generation (petascale) of life-science codes will be running in computing environments far more complex than those commonly used by biological researchers today. Furthermore, computational infrastructures will not appear without advance planning to make these systems easy to use and optimized for delivering a sustained hardware peak performance on biology applications with widely disparate computational requirements.

Biologists should embrace high-performance computing as a tool, and the computational infrastructure needs to occur at both the software and personnel levels. This can be facilitated by building a biological science network that connects computing and human resources for experiment, discovery, education, and teaching and ensures timely access and interactive teamwork-driven problem-solving. There have been only limited amounts of such integration in the past, but ISBE will be successful only if much more attention is paid to considerations including the following:

- Integration of modern enabling technologies with legacy and developing biological applications.
- Collaboration between computational scientists and biologists on how best to exploit and utilize high-performance computing resources.

The ISBE Data Generation Centres provide the necessary framework by allowing researchers from all related disciplines to work under the same roof to model complex biological systems in healthy and pathophysiological situations.