Infrastructure for Systems Biology Europe

Work Package 4: Data Generation Centres

Deliverable No: 4.1

Policy Document on vision, scope and structure of ISBE Data Generation Centres

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

Country: 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	D4.1
Deliverable title	Policy Document on vision, scope and structure of ISBE Data Generation Centres
Dissemination Level	PU
Status & version	Final
Number of pages	16
WP relevant to deliverable	WP4
Lead Participant	BRFAA, Dimitris Thanos
Author(s)	S.Kolyva, I.Michalopoulos,
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

Most biological processes governing health or disease (metabolic, sport, developmental, cancer, cardiovascular, neurodegenerative *etc.*) involve complex network interactions between hundreds of genes and proteins. Invariably, the complexity is enormous and every case becomes different, necessitating the integration of experimental quantitative data on a systems-wide level to obtain information about the state, dynamics and variability of living cells, organs, organisms and populations. The goal is to standardize these approaches and integrate them into predictive models. This phenomenon pertains to the understanding and promotion of health and the retardation of ageing, as well as to better diagnosis (e.g. through the development of novel biomarking strategies) and therapies (e.g. using new network targeting drugs) of disease. Where the latter is directly relevant to red biotechnology, the same issues apply to nutrition and white biotechnology, to green biotechology, as well as to the fields of bio-energy and ecology. Systems Biology integrates high-throughput technologies, model systems, molecular biology, biochemistry, engineering, information technologies, bioinformatics, clinical research and innovative engineering to understand how biological function emerges from interacting biological components and predict biological outcomes. This integration can only be achieved through a certain critical mass of experimentation, such as in genomics, and with the help of mathematical analyses, modelling, informatics and statistics. Biological networks, both intracellular and in-and-between whole cells, tissues and organisms, connect thousands of molecular and higher-order functions, such that functioning of any part of the network depends on different, remote parts. 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. It is important to emphasize that these data must be collected under physiological conditions and in a standardized manner in order to be useful to many investigators. Handling and interpreting these diverse data sets demands the use of a variety of computational, mathematical and statistical modelling techniques and can only be achieved with a critical mass in both the experimental and quantitative sciences.

To make this effort both effective and efficient requires the creation of an environment, an infrastructure, for systems biology for Europe that is capable of coordinating and leveraging the science base. This is the aim of the Infrastructure for Systems Biology in Europe (ISBE). ISBE's vision is to create a pan-European infrastructure for systems biology that will allow life scientists to understand the function of living organisms to a

much higher precision and in a holistic and predictive way. This allows intervention in the functioning of biological systems in a predictive and rational manner. The infrastructure enables scientists in academia, the health sector and industry to access and exploit the full potential of data-driven computational modelling of complex biological systems with the required reproducibility and validation. It provides the expertise, tools and resources to address current and future grand challenges in healthcare, agriculture and industrial biotechnology, thereby enhancing the wealth and well-being of the European citizens. ISBE will collect, develop, maintain and make available a wide range of tools and resources for the construction of maps of biological networks and predictive computational models of complex biological systems. This must be combined with expertise to analyse the quality of maps and models and to do model simulation and model validation. Algorithms and workflows used for system analysis will be transformed to toolboxes and software packages that are made widely available. ISBE will provide consultancy services with modelling experts for planning experiments for model construction; pairing of data experts and modellers for collaboration and construction of models from experimental data.

Systems biology models require data that are fit for modeling, both when models are built and for their validation. More often than not, disparate data sets must be combined and integrated (e.g. microscopy and proteomics and metabolomics) posing severe constraints on experimental and analytical procedures. This includes issues such as precision of data, sample preparation, data acquisition protocols and data analysis methodologies. To support the life sciences community ISBE will provide access through its National System Biology Centres (nSBCs) to services of data generation facilities that are able to provide model compliant data. Either such facilities may be integrated into the nSBC, or the nSBC may liaise with such facilities provided by other infrastructures, projects or initiatives.

Data Generation Centres

Internationally the life sciences undergo fundamental changes by transforming from highly dispersed and relatively low-cost research activities to big science; big in the sense of formation of large research consortia, huge data sets and big technologies, requiring major investments in hardware and expertise. More so, these technologies develop rapidly, requiring frequent updating and rapid depreciation. This holds for the Omics technologies - next generation sequencing, proteomics and metabolomics - and modern microscopy, as well as for infrastructures dealing with ‘big data’, i.e. bioinformatics, systems biology and e-science. These developments have an impact on the way research is organized, because no single institution – be it a university, national

research institute, or industry – is able to adequately cover all key technologies. Nonetheless, availability and easy access to these technologies is essential for research and development in health, agriculture, nutrition and biotechnology and therefore are essential for European innovative power and competitiveness. Present technology investment schemes are inadequate for building and maintaining an accessible and cost-effective high-end expertise and hardware infrastructure in the life sciences and unable to deal with the exponentially growing data volumes. This calls for a European life sciences technology agenda in which stakeholders join forces, including universities, national research institutes and funding agencies and industry.

ISBE is an initiative that aims to develop life sciences technology agenda and establish a consortium of stakeholders that together builds and manages a European high-end technology infrastructure that is accessible for the life sciences in academia and industry, offering state-of-the-art expertise, equipment and big data infrastructure and data stewardship.

The 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. Some clusters will focus on distinct conceptual aspects of biology - such as model organisms, model cell populations, diseases, biotechnology, ecology etc; others will have a central focus on the development and application of new 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. Importantly, ISBE will take a leading role in exploring, developing and establishing the necessary standards for experimentation and modelling in systems biology, critical for ensuring the delivery of reliable and consistent data across the infrastructure. A key deliverable of ISBE will be to bring focus to these disparate efforts by identifying, structuring and supporting large-scale research projects on major areas of urgent need in medicine, the biosciences and the economy - for example, human physiology and ageing, complex diseases, bioenergy and bio-manufacturing. Finally, such a focused European infrastructure will offer a single point of contact for access to a

network of best 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. The services presented in the ISBE Business Plan have been designed to address the challenges identified in the ISBE stakeholder engagement process. ISBE will offer services and resources that are useful for all branches of the life sciences, independent from the type or organism of biological system studied. Systems biology creates a remarkably strong unifying and theory-based foundaton in the life sciences.

Aims of the ISBE Data Generation Centres

Systems approaches require the collection, integration and storage of large data sets from genomics, proteomics, metabolomics and other –omic fields as well as molecular imaging with the goal to model life processes. The aim of the Data Generation work package (WP4) is to develop the appropriate institutional infrastructure and scientific mentality to support the collection of high throughput quantitative data in the biomedical and biotechnology fields in a way that is fit for systems biology. The overall objective of WP4 is to translate the strategic vision developed in WP3 and WP7 into an infrastructure plan for the construction phase of ISBE.

The aims of the Data Generation WP4 are:

- To document the existing physical infrastructures providing high throughput data generation of partner institutes in order to identify those that are relevant to ISBE.
- To develop a long-term, strategic vision of the role of high throughput data generation centres for systems biology research.
- To develop a distributed infrastructure prioritised plan to renovate existing infrastructures and build new infrastructures. To determine their contribution and impact on ISBE by constructing a roadmap of integrating relevant proposals for national and pan-European construction plans.
- To design and define standards and harmonise procedures in the operation and management of the data generation centres and to ensure implementation of harmonised and standardised operating practices among users.
- To define the needs of the European scientific community regarding access to the Data Generation Centres.
- To develop a plan for the implementation of European access and to develop the strategies that will establish the rules for providing European access to the Data Generation Centres of ISBE.

- To assess the needs, survey existing solutions and define the future strategy for developing a distributed data storage infrastructure for systems biology that can efficiently cope with unprecedented data volume and will be linked closely to the data generation centres and modelling hubs of ISBE (in cooperation with WP3).

Defining the goals of the Data Generation Centres

The goal of WP4 is to provide an integrated infrastructure to allow European scientists to collect high throughput biological data and to verify model systems using state-of-the-art technologies, including high throughput genomics and transcriptomics (DNA array systems, highly parallel DNA sequencers and advanced multiplexing technologies), advanced proteomics, metabolomics, high throughput imaging systems including microscopy, flow cytometry, automated cell analysis, etc. The ISBE Data Generation Centres will build on existing resources (physical infrastructure including buildings, large scale equipment, human resources, scientific and management expertise) distributed among partner institutes. These resources will provide the basis to develop a state-of-the-art distributed infrastructure to satisfy the needs of the European research and biotechnological communities for systems biology resources. However, despite the wealth of research facilities only a small percentage of the available data are good enough for systems biology work. Most of the available data available are not harmonized (lack of standardization) neither they have been collected under physiological conditions. Thus, there is an urgent need for the collection of high throughput data suitable for systems biology work.

A first approach has been made to define the outline of the Data Generation Centres. We adopted a three-pronged strategy by focusing on:

1. critical computational challenges arising from the development of new technologies – next-generation DNA sequencing, advanced proteomics, metabolomics, imaging, single cell technologies etc.
2. fundamental and long-term research pursuits – systems biology driven by heterogeneous data-intensive approaches, and
3. development of bioinformatics methods for advancing technologies

Twenty-first century biology seeks to integrate scientific understanding at multiple levels of biological abstraction, and it is holistic in the sense that it seeks an integrated understanding of biological systems through studying the set of interactions between components. Most existing biological data sets are unsuitable for systems biology modelling: they are incomplete, unannotated, or have been acquired for other

purposes. 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.

Because such an enormous, data-intensive effort is necessarily and inherently distributed over multiple laboratories and investigators, an infrastructure is necessary that facilitates the integration of experimental data, enables collaboration, and promotes communication among the various actors involved. Unique for modern systems biology however, is the fact that all these approaches are required simultaneously and interactively. The challenge is how to get all of the required expertise together. Centres for systems biology capable of supporting some of the many required expertise provides a solution to this question. The primary value of such an infrastructure resides in what it enables with respect to data management and analysis. Thus, in a biological context, machine-readable terminologies, vocabularies, ontologies, and structured grammars for constructing biological sentences are all necessary higher-level components of the infrastructure, as tools to help manage and analyze data. When coordination is difficult, researchers in different fields and at different sites tend to adopt different formats and representations of key information. As a result, their reconciliation or combination becomes difficult to achieve—and hence disciplinary (or subdisciplinary) boundaries become more difficult to break down. Without systematic archiving and curation of intermediate research results, useful data and information are often lost. Without common building blocks, research groups build their own application and middleware software, leading to wasted effort and time.

Services provided by the ISBE Data Generation Centres.

Europe houses a large number of Institutes with infrastructures capable of supporting the reliable collection of Systems Biology-grade data. The geographic distribution of these Institutes is unbalanced. As has been shown in Deliverable 4.2, more than 90% of the institutes are located in Northern Europe. These research Centres are operational, giving European investigators, predominantly from Northern Europe, access to an interlinked set of high-end key technologies (proteomics, metabolomics, next generation sequencing and bioimaging) and to data expertise and infrastructure: bioinformatics, systems biology, e-science). These facilities are tightly linked to research groups that have an international reputation in their technology field (see Deliverable

4.2). ISBE Data Generation Centres (DGCs) propose to take responsibility, on behalf of the consortium of stakeholders, for the development of technology hubs (personnel, hardware, consumables, lab-space, accessibility and periodic updating of equipment). Evident advantages for organizations that house a DGC are the easy access to advanced technologies, direct connection to ISBE and periodic updating of equipment and expertise.

Leading systems biology centres in most European member and associated states have been identified during ISBE's Preparatory Phase. Within each of the national communities, ISBE has pinpointed the leading centres and addressed a number of them to audit their capacity and interest in providing systems biology services and resources to a broad audience. Prominent researchers in more than 40 institutions across Europe have been consulted, and more than 80 available systems biology resources have been identified with this exercise Discussions with potential nSBCs that could provide access to Data Generation Centres will be initiated, aiming at identifying candidate nSBCs/DGCs. A transparent and open nomination process will be started later in 2015 in the Preparatory Phase and continued during the Interim Phase. It will require candidate nSBCs/DGCs to seek financial support from a national funder together with endorsement by the host institution. Delivery of existing resources and services via national institutions will be based on their relevant national and EU procurement regulations. Further services will be delivered in a consistent manner, having been identified through involvement of national funders through the strategic plan for developing ISBE. Presently, a number of prominent systems biology institutions and research groups have been audited with respect to their expertise and willingness to become a provider as part of a national Systems Biology Centre (nSBC) of ISBE. More than 40 institutions across (9) countries have provided detailed information regarding their infrastructure capabilities and recognize the value of offering their expertise and services to a broader community of users via a pan-European infrastructure, and thus manifested interest for being part of ISBE as providers.

Taken together, the ISBE DGCs will provide a range of unique expertise or services that will contribute to delivering the continuity and positive impact to their national communities, whilst ensuring that ISBEs operates as an effective and cohesive provider giving European added-value.

Below are the technical expertise and the type of data to be provided by the ISBE DGCs.

• DNA Sequencing: Leading-edge molecular biology and sequencing capabilities at multi-petabase scale (genomic sequencing, RNAseq, miRNAseq, Chip-seq etc)

- DNA microarray technologies: (Copy number variant, Gene Expression analysis, Genotyping analysis, DNA methylation etc) are some of the technologies in high demand by European scientists.

- Advanced proteomics: Mass spectrometry based techniques have become widely available through out European Institutions, MS-peptide and protein identification, Quantitative MS, MS posttranslational modifications, Protein and peptide arrays, Antibody arrays and 2- Dimensional electrophoresis for proteomics are available in the majority of Systems Biology institutions.

- Metabolomics technologies: Targeted quantitative metabolomics, mass spectrometry, on-targeted metabolomics, plant and microbial metabolomics, high throughput metabolomics, Clinical metabolomics are metabolomics technologies available in European Institutes.

- Single cell technologies: Able to interrogate at a genomic, transcriptomic, proteomic and metabolomic level of thousands of isolated single cells from every organism. Single cell approaches are increasingly receiving important attention for proteomics, transcriptomics and functional genomics. A large number of systems biology scientists appreciate the significance of single cell DNA and RNA sequencing. The general consensus among the scientists participated in our interviews is that next generation sequencing is not yet at the peak of its usage, and that in the near-future the main technological developments will also involve Next gen sequencing but focusing to develop library technologies to read longer pieces of DNA, to improve the accuracy of sequencing and to capture as many transcripts as possible from single cells.

- Imaging Technologies: Light microscopy, Advanced Light microscopy, Electron microscopy, Probe microscopy, Correlative light and electron, PET, SPECT, MRI, CT, Ultrasound, Optical tomography are available in European institutions. In microscopy there are couple of techniques that might have taken the revolutionary step (as 2PPM) and are currently in the fermentation phase.

• High-Throughput Functional Annotation of Genomes: Conversion of sequence data into biological insights

• Massive Scale Sample Preparation: Development of technologies for designing and carrying out complex “grand challenge” projects involving hundreds of collaborators, thousands of conditions, and tens of thousands of samples

• Synthesis: Able to synthesize and express thousands of genes and large pieces of DNA to engineer complex organisms for hypothesis testing.

Existing high throughput data generation centres

Systems Biology is an integrated experimental, informational, and computational science. It has benefited from advances in fields such as genomics, proteomics, metabolomics, and others that employ high-throughput technologies and is driven by innovations in mathematics, computational analysis and simulation. Biologists may now pay more attention to understanding how biological components work together to produce system behaviors rather than focusing exclusively on the properties of individual molecules and pathways, although the latter is foundational for such inquiries. The adoption of a systems approach is providing new knowledge in many areas of Life Sciences and biomedical research including cell dynamics and signaling networks, global metabolic fluxes, and responses to drugs (and guidance in their development). It is expected that new, fundamental rules governing systems behavior at various organizational levels – and how these levels are integrated - will emerge from these studies. However, there continue to be significant conceptual, technological, and cultural challenges to the realization of the systems biology goals. It is the purpose of this initiative to promote innovative responses to these challenges.

Systems biology is an interdisciplinary science that derives from biology, mathematics, computer science, physics, engineering, and other disciplines. The infusion of theories and techniques from other fields and their integration is establishing new methodologies for problem definition, hypothesis generation and testing, and experimental approaches in biomedical science. Most biological systems are too complex for even the most powerful computational models to capture all system properties. A useful model, however, should conceptualize and formalize the system under study such that it becomes a powerful hypothesis generator. To accomplish this, a certain level of abstraction may be required that focuses on the system properties of interest while neglecting some of the other details. In this regard, there is a need for additional research on such To accomplish this, a certain level of abstraction may be required that focuses on the system properties of interest while neglecting some of the other details. In this regard, there is a need for additional research on such issues as system modulation, parameter estimation and optimization, and model scalability, with the goal of learning how models can be usefully employed to understand and predict

biological behavior. An attractive way to achieve this goal is to develop collaborations between biologists and experts from other fields. Such interdisciplinary collaborations will likely provide the inspiration for the generation of new conceptual thinking, as well as new systems biologists.

Currently, technological deficits exist in mathematics, computation and experimentation. These include lack of standards and quality control measures in data collection and software engineering. Growing volumes of data from diverse high-throughput experiments provide unprecedented opportunities for computational biologists. However, a high level of heterogeneity in data quality and experimental conditions hampers data comparison, integration, and application in computational modeling. Similar issues exist in software development. The lack of software engineering standards and sufficient documentation has limited software re-use, resulting in unnecessary duplication of efforts, and difficulty in comparing one program to another. Experimentally, there is a demand for the development of novel (and low-cost) methodologies to miniaturize, standardize, and automate high-throughput data collection such that computational models can be populated with data specifically selected for tests of those models. In some cases, measurements from single cells are of great value. 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.

Structure of DGCs

The *coordinating Systems Biology Centre* (cSBC) will maintain the overall overview of the scientific and technical capabilities of all nSBCs and their associated DGCs and coordinate their services, as well as identifying gaps in expertise together with the Central ISBE Office (CIO). In addition, the cSBC will manage complex or extensive user requests that require cooperation of two or more nSBCs working together to provide the requested services. It will support pan-European community activities including training and development of community standards. Managing individual DGCs within the context of ISBE will be the responsibility of nSBCs and the national authorities. The services offered by nSBCs will reflect their national priorities, and so service provision will be heterogeneously distributed across countries. However, this heterogeneity will as far as possible be hidden from end users, thanks to the use of common standards, design of coherent interfaces, and coordination from the cSBC so that all nSBCs are aware of what the others have to offer.

The staffing and funding of nSBCs remains for ISBE members to decide, but a tentative proposal might be:

• A full-time paid coordinator.

• Highly competent engineers/technicians to run a national help desk.

• Resources to attend conferences and to interact closely with other systems biology centres.

• Resources to participate in large-scale and prestigious European projects

Open- innovation environment

One of the key objectives of DGCs is to provide a catalyst for the establishment of integrated technology programs and concrete valorisation opportunities. Here, DGCs will establish project development strategies that supports open innovation, together with a transparent and uniform IP policy. Meetings, workshops and other events will be organised, aiming to make DGCs a platform for match making between researchers and industry, to drive joint vision development towards marketable applications.

As the driver of European life sciences technology agenda, DGCs aim to align and synergise investments in expertise and infrastructure that are made by its stakeholders, and universities/Institutes. It is important to note that DGCs bundle existing infrastructures, expertise and partnerships, rather than starting new Centres. In terms of overall investments, establishing DGCs should reduce costs and increases both quality and efficiency. Access fees to DGCs facilities in principle should be covered by research grants or other funding mechanisms. Costs of technologies should therefore be explicitly budgeted in grant applications. Access to DGCs by commercial parties will be based on market-level fees, with benefits for DGCs partners.

International embedding

A European life sciences infrastructure such as ISBE cannot be developed without embedding it in the variety of large international, in particular European, initiatives. The international dimension significantly enhances the scope of DGCs and gives the European life sciences access to international infrastructural facilities. DGCs therefore are the natural node in the ESFRI hubs-and-spoke networks of European expertise centres in bioinformatics (ESFRI-ELIXIR), advanced bioimaging (ESFRI-EuroBioImanging) Biobanking (ESFRI-BBMRI) and translational research (ESFRI-EATRIS).