Proposal Full Title:

From Data To Models:

New Bioinformatics Methods and Tools for Data-Driven, Predictive Dynamic Modelling in Biotechnological Applications

Proposal Acronym: **BioPreDyn**

Type of funding scheme: **Collaborative Project**
(small or medium-scale focused research project targeted to SMEs)

Call identifier: **FP7-KBBE-2011-5**

Work programme topic addressed:

KBBE.2011.3.6-01:
Increasing the accessibility, usability and predictive capacities of bioinformatics tools for biotechnology applications.

Date of preparation: **January 24th, 2011**

Name of the co-ordinating person: **Johannes Jaeger**

Coordinating organisation: **Centre de Regulació Genòmica (CRG), Barcelona, Spain**

E-mail: yogi.jaeger@crg.es
Telephone: +34 93 316 02 85
Fax: +34 93 396 99 83

List of participants:

Participant No.	Participant Organisation Name	Short Name	Team Leader	Country
1	Centre de Regulació Genòmica	CRG	Johannes Jaeger	ES
2	Agencia Estatal Consejo Superior de Investigaciones Científicas	CSIC	Julio R. Banga	ES
3	European Molecular Biology Laboratory	EMBL	Julio Saez-Rodriguez	DE
4	Universiteit van Amsterdam	UvA	Jaap A. Kaandorp	NL
5	Centrum Wiskunde & Informatica	CWI	Joke Blom	NL
6	Telethon Institute of Genetics and Medicine	FTELE.IGM	Diego di Bernardo	IT
7	University of Manchester	UNIMAN	Pedro Mendes	UK
8	University of Sheffield	USheff	Neil Lawrence Magnus Rattray	UK
9	Complex Systems Modeling (CoSMo)	CSM	Eric Boix	FR
10	Insilico Biotechnology	INSIL	Klaus Mauch	DE
11	Fluxome	FS	Jochen Förster	DK

Table of Contents

Content

Proposal

Summary

Currently, biologists are collecting enormous amounts of 'omics' data in a vast number of different databases. Predictive, data-driven computational models are needed to understand the complex, multi-scale biological networks underlying these high-throughput datasets. Such models are non-linear and contain many parameters, which are difficult (or impossible) to measure directly. Instead, parameters need to be inferred from data. This approach is called reverse-engineering. It has tremendous potential for several areas, such as biotechnology and systems biology, since it allows us to develop models with unprecedented accuracy and predictive power. This is achieved through an iterative refinement of our models compared to quantitative 'omics' data, a process called the systems-biology modelling cycle. Many methods have been developed that deal with specific steps in this cycle (data analysis, model building/discrimination, parameter estimation/identifiability analysis, uncertainty quantification, and optimal experimental design), but we still lack an over-arching, easy-to-use software framework that supports the modelling cycle in its entirety, allowing its widespread application. This project aims at improving accessibility of the data, and developing novel algorithms and tools implemented in such a general framework, which will enable the efficient transfer of cutting-edge modelling and optimisation methods from an academic research setting to private biotechnology partners. We will use representative biological and biotechnological applications as benchmark problems to develop robust and generally applicable methodology. The availability of such tools to the biotechnology sector (and other industries) will greatly enhance our ability to design and optimise complex production processes, especially those of nutraceuticals, biopharmaceuticals, or fine chemicals based on engineered organisms such as bacteria, yeast or plants.

1: Scientific & Technical Quality, Relevant to the Topics of the Call

1.1 Concept and Objectives

Mathematical and computational **models** (used in conjunction with quantitative data) are central in bioinformatics and systems biology. Models provide **new ways to exploit and interpret existing datasets**, **generate novel and testable hypotheses**, and enable us to gain a **mechanistic understanding** of the function of **complex biological systems**. They also support a **quantitative framework for interventions** involved in the health and biotechnological sectors. A particularly interesting application is the design and optimisation of biotechnological production processes based on engineered microbial systems, cell lines, and (soon) synthetic biology.

Since the amount and quality of experimental "omics" data continues to increase rapidly, we are in **great need of implementing integration and exploitation of the data, and developing methods for rigorous and systematic model building, validation, and analysis**, which can handle this **complexity**. Such methods are currently being developed by multiple academic research groups, but their wider application—especially in an industrial biotechnology context—is seriously hampered by the lack of standardisation and powerful, easy-to-use, reliable software tools. This project aims at resolving this issue, by bringing together academic labs that manage large databases and develop cutting-edge model-building, analysis, and optimisation algorithms with small and medium enterprises (SMEs) that can implement these tools in a consistent, and well-supported software framework and apply them to biotechnological applications. Our planned collaboration between algorithm developers and biotechnology companies will facilitate the transfer of information and code from an academic setting to commercial application, and will thereby strengthen European competitiveness in the fields of systems/synthetic biology and biotechnological production processes based on engineered biological systems.

1.1.1 Modelling of Biological Systems

Models are the central elements in hypothesis-driven research in systems biology. A model represents a computable set of assumptions and hypotheses—encoded explicitly and quantitatively by rules and equations—that need to be tested or supported experimentally (Kitano, 2002).

Complex biological systems are usually represented by **networks** (also called graphs). A network is an abstraction of a complex system that is extremely useful—when used in the proper way—to understand and predict the system's behaviour. In a network, the system is divided into components; each component is abstracted as a single **node**, and **edges** between pairs of nodes represent (dynamic) interactions.

An edge can either represent a direct physical interaction—the basis of **mechanistic models** (e.g. in gene regulatory networks, a transcription factor regulating its transcriptional target, or a kinase phosphorylating its substrate)—or influence interactions—the basis of **phenomenological models** (i.e. an enzyme whose concentration changes the quantity of a metabolite, which in turn affects the level of another protein).

There are two main difficulties with modelling complex biological systems:

(1) **The choice of scale and scope of the model**. Stelling (2004) argues that mechanistic dynamical modelling is the most obvious candidate for achieving a system-wide understanding of biological systems. But scaling of such models to the whole-genome level is not easily achieved, while modelling molecular details is not always possible (nor always desirable). Therefore, it is extremely important to find the right compromise, that is, to choose the adequate scope and level of detail for a model. This compromise needs to be firmly and systematically grounded in the available preliminary evidence, and the research question at hand.

(2) **The choice of phenomenological modelling framework**. Many important processes, such as eukaryotic transcriptional regulation, are not yet understood in molecular detail. Therefore, models need to approximate them at the phenomenological level such that the interactions among system variables are defined in an operational rather than a mechanistic way (Wolkenhauer & Mesarovic 2005). The problem is that there are many reasonable phenomenological modelling frameworks available. In most cases, it is not evident which alternative framework is best suited for a given problem.

For these reasons, there is a clear need for sound and robust procedures to build mathematical and computational models of biological systems from the vast amount of data generated from the different 'omics' disciplines today. One issue here is **accessibility, standardisation and integration of large, heterogeneous datasets**. Another is **system identification**, a key area in systems engineering, which deals with the development of mathematical models of dynamic systems from specific input/output datasets (Ljung 1999; Walter & Pronzato 1997). The modelling itself requires advanced techniques for multi-scale/hierarchical simulation, rigorous model validation/comparison, and uncertainty quantification.

A third important aspect is developing a rigorous protocol for the systems biology **modelling cycle** (Fig. 1), which addresses all possible sources of errors in the cycle. Whereas a lab protocol is common practice for "wet" experiments, it is not so for the "dry" part. Systems biology requires a protocol for the complete cycle to integrate the whole biological knowledge discovery process. Experimental **data analysis** is not just producing clean datasets but also information about the experimental error to be used in distance measures and re-sampling strategies for **rigorous validation** (see below). Exploratory data analysis **integrated** with **data** from the literature extracts biological knowledge leading to a number of hypotheses that together with the assumptions are formulated into mathematical models. Extra assumptions will be required, like a choice of scale in time, space and chemical and molecular detail. To test these assumptions **multi-scale models** and coarse-graining techniques will be necessary. The numerical implementation of all models has to be verified for both parameter and state space (**model verification**). Using **system identification** the parameters (with uncertainties) of the model can be inferred from the data using adequate distance measures. System analysis addresses the propagation of uncertainties in parameters and state-variables in the model-results and validates the models with unseen data (**uncertainty quantification** and **model validation**). The surviving models can be trusted to reflect the hypotheses and a new cycle can be started by **optimal experimental design** to **discriminate** between alternative models or to obtain more data to improve existing models. State-of-the-art methods (indicated by blue colour in Fig. 1) will be developed, and will be combined with methods from fields yet to be explored (red in Fig. 1) into a rigorous protocol, which will be validated and tested.

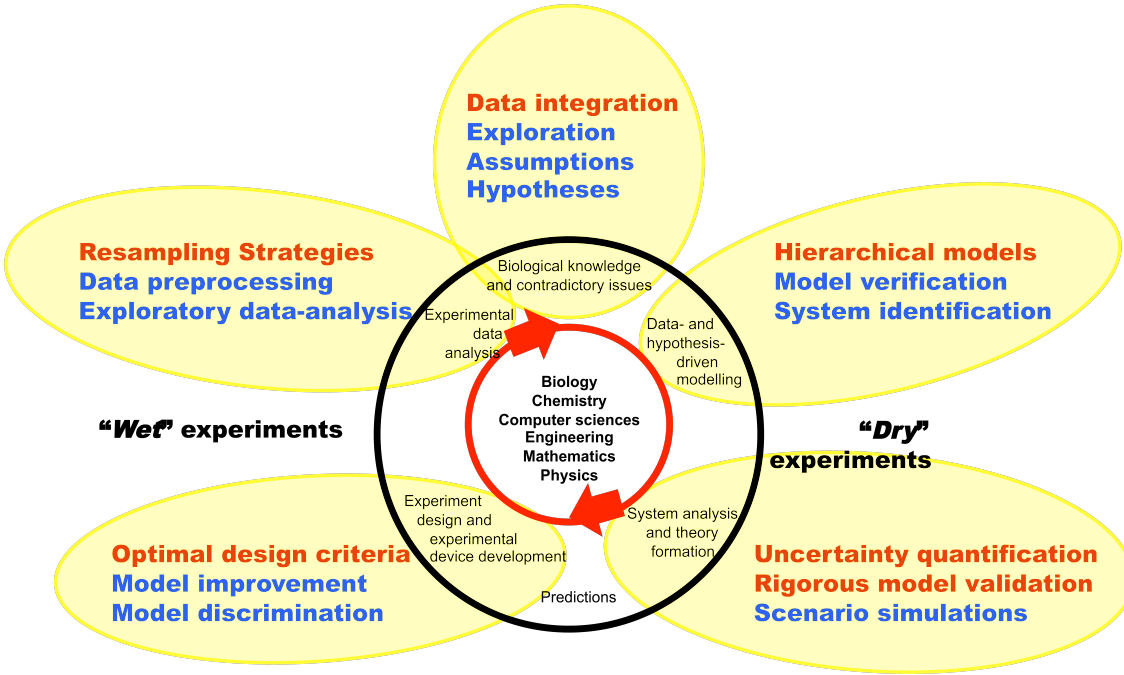

Figure 1. **The Systems Biology Modelling Cycle.** Blue indicates existing methods, red indicates innovative research (adapted from Kitano 2002).

1.1.2. Reverse-Engineering of Biological Networks

The fact that most models dealing with complex biological systems are of a phenomeno-logical nature implies that many model parameters cannot be measured directly. For example, connectionist gene network models (Mjolsness *et al.* 1991; Jaeger *et al.* 2004a) represent each regulatory interaction in a system by a single number (positive for activation, negative for repression), simplifying the complex molecular reality of transcription factor binding sites, enhancers, silencers, insulators and chromatin structure that determine the regulatory nature of a molecular interaction *in vivo*. Therefore, there is no straightforward connection between such a summary regulatory weight, and any measurable biophysical quantity (e.g. the dissociation constant, or fractional occupancy of a transcription factor at its binding site).

Parameters that cannot be measured need to be inferred. This approach, called **reverse-engineering** of biological systems, can be defined as the process of identifying regulatory interactions from experimental data through computational analysis. Gene expression data from microarrays (or more recently, RNA-seq) are typically used for this purpose (see, for example, De Smet & Marchal 2010). Microarrays provide quantitative expression data for a large number of genes, which is obtained by hybridizing extracted total RNA to oligonucleotide probes on the array, representing an integrated measurement of the state of cells in a tissue under specific conditions over time. Similarly, RNA-seq provides quantitative expression data for a large number of genes through deep sequencing of the extracted total RNA. Both microarrays and RNA-seq have the disadvantage that it is difficult to measure spatially specific expression patterns, which are important for problems in developmental biology, but also in genetic engineering and synthetic biology in animals and plants. In a number of cases, gene networks have been inferred from spatial gene expression patterns based on detection of mRNAs or proteins in living or fixed tissues or embryos (Jaeger *et al.* 2004a).

There are three main approaches that have been successfully applied to reverse-engineer metabolic, signalling, and gene regulatory networks (Bansal *et al.* 2007): the Bayesian Network (BN) approach, an approach based on information theory (mutual information—MI), and approaches based on differential equation (DE) models. Both BN and MI approaches are computationally efficient, and relatively easily scalable to large gene networks, but in general only allow us to obtain a topological (i.e. static) map of gene-gene interactions from the experimental data. Approaches based on DEs, on the other hand, aim at identifying a dynamical model of the underlying network, in addition to the identification of the static network map. Such models can be used to simulate network dynamics *in silico*. However, DE-based methods are computationally expensive and do not yet scale well beyond networks containing a relatively modest number of genes.

Our main focus in this project is on reverse-engineering approaches based on DE models (although we will also use Bayesian inference and approaches based on mutual information where suitable, see below). Such approaches consist of four basic steps (see, for example, Reinitz & Sharp 1995; Jaeger *et al.* 2004a): (1) A suitable quantitative dataset is generated, which measures a combination of state variables (for example, mRNA or protein concentrations) of the system. (2) A general model—based on ordinary (ODEs) or partial differential equations (PDEs)—is formulated. (3) The model is fit to the data by means of global non-linear optimisation. Thereby, the model is solved numerically, and model parameters are altered while selecting solutions that resemble the data increasingly closely, until the model reproduces the data faithfully and reliably. (4) Biological insight is gained, and predictions are derived, by analyzing the dynamical behaviour and the parameter values of the solution. In this way, dynamical models are used as computational tools to extract regulatory information from data.

1.1.3 The Model-Building Cycle

Model building is an **iterative process**, usually represented as a **cycle** (Fig. 1). It starts from the definition of the purpose of the model. In other words, modelling must start with a specific question to be addressed, often induced by **data analysis** and knowledge from literature. This question conditions the **selection of the modelling framework**: Which components, and which processes should be included? Which levels of detail (molecular, cellular, tissue-level, organismic) should be considered? Which processes can be approximated (and in which way)? Which ones need to be

modelled in molecular detail? Once these questions are clarified, a modelling framework is chosen and a first mathematical model is proposed taking into account available a priori knowledge and preliminary experimental data. Often, one model will cover multiple hierarchical levels of detail (**multi-scale modelling**).

Such preliminary models usually contain unknown parameters, which are difficult or even impossible to measure. These parameters must therefore be estimated by means of fits to experimental data (**reverse-engineering**). This process is called optimisation or parameter estimation (Ashyraliyev *et al.* 2009a). Most biological problems are highly complex and non-linear, such that model fitting is difficult and computationally expensive due to the large size and high dimensionality of parameter space as well as the presence of numerous local optimisation minima. Specialized, cutting-edge **global optimisation algorithms**, such as simulated annealing, evolutionary algorithms, or scatter search, are required to carry out precise, reliable and efficient global optimisation of network models (see, for example, Moles *et al.* 2003; Jaeger *et al.* 2004a,b; Perkins *et al.* 2006; Fomekong-Nanfack *et al.* 2007; Rodriguez-Fernandez *et al.* 2006b). In many cases, multiple optimisation criteria must be considered (goodness of fit, robustness or biological realism of the resulting mechanism, etc), and therefore, **multi-objective optimisation (MOO)** must be employed (Handl *et al.* 2007). The proper choice and implementation of optimisation criteria is the subject of a research field called **measure design** (Oberkampf & Barone 2006; Deb 2009).

We need to know whether it is possible to uniquely determine parameter values (**parameter identifiability analysis**). Ideally this is done before parameter estimation has been carried out (a priori identifiability), but this is often difficult to achive (Walter & Prozato 1997; Jaqaman & Danuser 2006). In these cases, parameter identifiability analysis is done after parameter estimation (a posteriori) (Gadkar *et al.* 2005; Balsa-Canto *et al.* 2010). This analysis not only uncovers which parameters are ill-determined, but also whether such parameter 'sloppiness' is due to insufficient data, or parameter correlations within the model (Gutenkunst *et al.* 2007; Ashyraliyev *et al.* 2008, 2009b). Finally, the model should be **validated**, i.e. a new or unused set of experimental data should be compared with the model simulations. Once this is done, the theory of **optimal experimental design (OED)** can be used to determine in which ways to expand our existing datasets to improve identifiability and uniqueness of optimisation solutions and to discriminate between rivalling models/hypotheses.

Bayesian methods provide an alternative paradigm for model inference, parameter estimation and optimal experimental design (Lawrence *et al.* 2010). In the **Bayesian framework** we compute a *posterior distribution* of models (or model parameters), which is a weighted ensemble of models consistent with the available data and prior domain knowledge. In this approach we do not restrict ourselves to a single "optimal" parameter or model but we identify a distribution, which captures the uncertainty of our parameter and model inference. Asymptotic Bayesian parameter inference (MAP learning) is very similar to the global optimization approach but non-asymptotic methods, e.g. those based on variational inference methods or Markov Chain Monte Carlo (MCMC) sampling, are quite different as they may retain a broad posterior distribution over models. The posterior distribution can be very naturally applied in the context of experimental design since we can select experiments that maximise the information gain by giving a large expected change to the posterior distribution, as quantified by some information theoretic divergence measure.

Preliminary models provide **predictions**, which must be (in)validated with new experi-ments, revealing in most cases a number of deficiencies. Consequently, a new model structure and/or a new (optimal) experimental design must be planned. **Model discrimi-nation** (also called **model selection**, not to be confused with the initial selection of the modelling framework described above) and **model ranking** methods are powerful tools that aid in the choice of alternative models (Vyshemirsky & Girolami 2008; Cedersund & Roll 2009). **Model reduction** can also be applied at this step, to simplify the optimisation procedure and/or the biological interpretation of results (Okino & Mavrovouniotis 1998; Radulescu *et al.* 2008). This process is repeated iteratively until the model is considered satisfactory.

1.1.4 Uncertainty Quantification

Unfortunately, all parts of the cycle contain errors and uncertainties that collectively affect the predictions: (i) It is not always possible to acquire the relevant experimental data or the

measurements contain uncertainties (systematic and random). (ii) The theoretical or mathematical model is not describing the reality (or more precisely, the quantities of interest) adequately. (iii) Simulating a mathematical model introduces numerical errors. And (iv), model parameters and initial conditions are not known (with sufficient precision) (Fig. 2) (for reviews on this subject see Karniakidis & Glimm, 2006; Oden *et al*. 2010a,b). The simplest approach is to quantify these errors separately. **Verification**—the error control of the numerical algorithms and the computational implementation—is often done when developing the algorithms but the resulting error estimates are mostly ignored or concealed. When **inferring the model parameters**, a probabilistic error estimate—assuming only known experimental errors—can be easily computed, but again these results are often not used in the subsequent steps of the cycle. Finally, the **validation** step is mostly neglected, sometimes due to a lack of experimental data, and if it is addressed there is no distinction made between validating the current model/parameter set with respect to the experimental data—re-sampling the dataset used for inferring and for validation—and validating the theoretical model e.g. by experimentally testing model predictions.

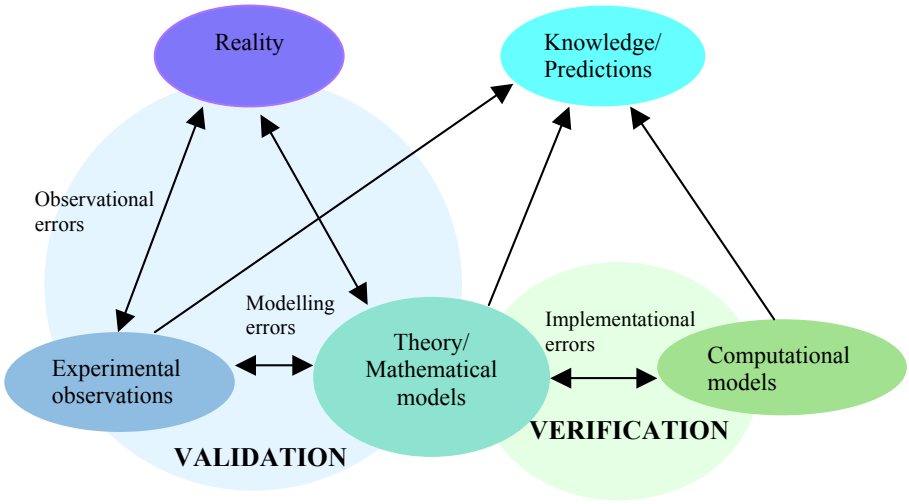

Figure 2. **Rigorous modelling needs to address all possible sources of errors** to establish their influence on the knowledge based on the experiments, the theoretical model and computational model simulations (adapted from Oden *et al.* 2010).

If the modelling cycle is put in a Bayesian framework, it is possible to link all errors in a probabilistic way and to discriminate between multiple heterogeneous models (Robert 2007; Vyshemirsky & Girolami 2008). In the Bayesian paradigm, probability distributions are used to describe data observation errors (stochastic) and model errors (uncertainty) in a consistent and well-defined way. This is an attractive unifying perspective, which we will pursue where possible, but it should be acknowledged there are very significant computational challenges when applying Bayesian methods over complex model spaces. An important focus of the Bayesian approach is therefore on the development of more efficient algorithms based on sampling (MCMC) and functional approximations (e.g. variational inference) (Lawrence *et al.* 2010).

1.1.5 Biological and Biotechnological Applications

The model building cycle described above can be applied to a **wide range of scientific problems and biotechnological applications**. Traditionally, academic research in the field has focussed on the study of metabolic, signalling or genetic networks involved in physiology or development (see, for example, Moles *et al.* 2003; Feng & Ratitz 2004; Jaeger *et al.* 2004a; Gadkar & Gunawan 2005, Honkela *et al.* 2010). In this context, the reverse-engineering approach is used to infer **regulatory interactions** among systems components, which explain the system's dynamical behaviour. Within our project, the academic partners will focus on such applications: microbial large-scale metabolic and transcriptional networks (in *S. cervisiae* and *E. coli*), cellular signalling networks (focussing on the Chinese Hamster Ovary, CHO, cell line, used for the production of eukaryote-specific products such as antibodies), and gene networks involved in biological pattern formation. These problems can be seen as benchmarks used to test and calibrate our methods.

The true potential of reverse-engineering and optimisation lies in their application to **industrial biotechnological processes**, which is one of the main aims of this project. This involves the modelling of metabolic and gene regulatory processes (as described above), where the parameters to be optimised are engineered regulatory interactions and processes involved in the production of nutraceutical ingredients (food additives) or other components. In other words, reverse-engineering methods will allow biotechnology companies to design and optimize their production processes in a much more reliable, predictive and quantitative way. The successful application of these methods will have a tremendous impact on the industry.

There are many more applications of reverse-engineering, which go beyond the scope of this project: for example, modelling complex proteomics datasets for diagnostics of complex diseases, prediction of chemical component candidates in rational drug design, complex sequence alignments arising in comparative genomics, or modelling of ecological networks to enable efficient resource management and protection policies. The methods developed in this proposal will be easily transferrable to a wide range of domains of application.

1.1.6 Need for Novel Methods

We have argued that modelling biological systems through reverse-engineering is a powerful and promising approach, with a large number of potential applications. However, this approach is still in its infancy. It has not yet been applied to many different systems, there is no general agreement yet which algorithms and tools are appropriate under which specific conditions, and there are no easy-to-use, integrated, cutting-edge software tools available for end users such as SMEs.

In light of this, there is a clear need for novel, powerful and integrated methods for reverse-engineering and biological modelling which are able to handle the special requirements that arise from complex biological datasets. In particular, our methods need

- to be able to deal with uncertainty or noise in data, or incomplete datasets,
- to enable effective integration and visualisation of databases and other heterogeneous data sources used for model development,
- to support diverse, multi-scale models and rigorous procedures for model identification and validation,
- to implement a diverse range of global, non-linear optimisation algorithms for parameter estimation, and to aid the user in choosing an appropriate cost function,
- to enable a priori and a posteriori parameter identifiability analysis,
- to allow us to implement optimal experimental designs for improving models based on evidence from parameter identifiability and uncertainty quantification,
- to support methods for model comparison and ranking to chose the most appropriate phenomenological modelling framework for a given problem,
- and finally, to implement powerful methods for computational analysis of the dynamical behaviour of a system, enabling us to gain biological insight from our models.

We aim at both developing improved methods and implementing them in a unified, user-friendly software framework. Since many of these methods are computationally very intensive, particular attention will be paid to the computational implementation of these tools for high-performance (parallel) computing (incl. GPU-based machines).

1.1.7 Objectives of the Project

BioPreDyn aims to develop **new bioinformatics methods and tools** for data-driven, **predictive dynamic modelling** in biological and biotechnological applications. The main objectives of BioPreDyn are structured in four groups, three vertical (methodological) objectives and one horizontal (applications) objective, as shown in Fig. 3.

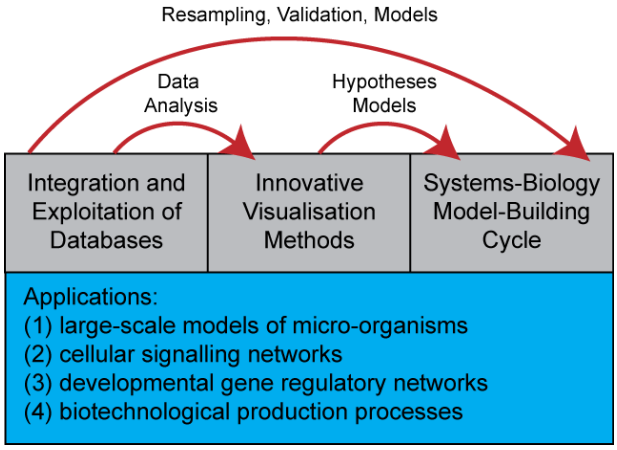

Figure 3: Main Objectives of this Project: three vertical (methodological) objectives and one horizontal (applications) objective.

The details of each objective and their relationship with the topics addressed by the call (***KBBE.2011.3.6-01*** *Increasing the accessibility, usability and predictive capacities of bioinformatics tools for biotechnology applications*) are as follows:

Objective 1: To develop tools for **integrating and exploiting databases**, especially those with dynamic expression data. The key novelty here is the development of methods and tools for handling of databases and other data sources containing time- (and space-) dependent biological data. This objective fits well with the call, since integration of databases is one of the challenges mentioned explicitly.

Objective 2: To implement **innovative visualisation methods** for data analysis and model development, with emphasis on dynamical models: as in the previous objective, a key novel aspect of our proposal is the consideration of biological data distributed over time and space. This objective also fits with the call, where innovative visualisation methods are highlighted as one of the main research themes.

Objective 3: To develop integrated software tools and workflows to support the **model building cycle:** currently there is a lack of tools for supporting the full cycle of dynamic modelling and reverse-engineering biological systems. In this project we will develop proper procedures and workflows for multi-scale model identification and building, measure design, parameter estimation by global non-linear optimisation, parameter identifiability analysis, model comparison, and optimal experimental design. This approach fits perfectly with another key topic of the call: the need for increased interpretative and predictive capacity of data, taking into account the complexity of living systems.

Objective 4: To apply these methods to a variety of **illustrative biotechnological and biological problems** in both academic and corporate settings: the new methods and tools to be developed in objectives 1–3 will be generally applicable. Their performance will be tested by considering several key biotechnological and biological applications:

(a) Large-scale dynamic modelling of metabolism and gene regulation in microorganisms (*Escherichia coli, Saccharomyces cerevisiae*) and eukaryotic cell lines.
(b) Cellular signalling networks with a special focus on the CHO cell line used for biotechnological production processes.
(c) Inference of developmental gene regulatory networks in fruit flies (*Drosophila)* and cnidarians (*Nematostella*).
(d) Mechanistic and comprehensive modelling of biotechnological production processes based on transgenic microorganisms.

1.2 Progress Beyond the State-Of-The-Art

1.2.1 Integration and Exploitation of Databases

A cell can be described as an ensemble of interacting biological entities (messenger and other RNAs, proteins, metabolites, organelles, compartments, etc). In multi-cellular organisms, cells themselves interact within and between various tissues and organs. The hierarchical, collective behaviour of all these entities underlies the observed phenotypes. Great effort is being put into research identifying and mapping networks of interactions among biomolecules in various biological systems, from microorganisms to vertebrates/humans. Massive amounts of heterogeneous data concerning the levels and regulatory interactions of network components have been, and are being, collected by laboratories world-wide using a variety of high-throughput experimental techniques. There are many types of interactions depending on the molecules being considered, and the function being studied:

Metabolic interactions were the first to be systematically mapped and, therefore, the field is quite advanced in prokaryotes and simple eukaryotes. We also have a good grasp of parts of the metabolic network in mammalian cells, which has been annotated and collected in public databases such as Reactome (www.reactome.org) and KEGG (www.genome.jp/kegg).

Then there are studies mapping **protein-protein interactions**, where each protein is seen as a node in a network, and two proteins are connected by an edge if they are part of the same protein complex. Human protein interactions, for example, have been mapped using the yeast-two-hybrid technique (Y2H), and automated data mining of the literature. The HPRD database (www.hprd.org) contains more than 38,000 such interactions.

Many studies have mapped **transcriptional regulatory interactions**. For example, chromatin immuno-precipitation techniques followed by microarray hybridisation (ChIP-Chip), or deep-sequencing (ChIP-seq), have enabled the identification of transcriptional interactions among all known yeast transcription factors (Lee *et al.* 2002). Similar efforts are being undertaken in model organisms such as *Drosophila* (Li *et al.* 2008). The translation of these techniques to mammalian cells has proven to be more difficult, because it is not trivial to map binding sites to the genes that are regulated by them. New experimental techniques have been proposed for identifying enhancer-promoter interactions (5C, Chia-PET) and may help solve this enhancer assignment problem.

The recent discovery of microRNAs (miRNAs) and their biological functions has generated a global effort in identifying **microRNA targets**. Several targets of miRNAs in human have been identified both experimentally, and by sequence/expression analysis. These have been collected and annotated in databases such as miRBase (microrna.sanger.ac.uk).

In addition, recent efforts have begun to identify all of the human **protein kinases** and their **phosophrylation** targets. Machine learning techniques have been successfully applied to identify kinase targets. These predictions are available in public databases (e.g. networkin.info).

Large datasets describing gene expression profiles are available from databases, such as ArrayExpress (www.ebi.ac.uk/arrayexpress), the Gene Expression Omnibus (GEO; www.ncbi.nlm.nih.gov/geo), or NCBI's short-read archive (SRA; www.ncbi.nlm.nih.gov/sra; for RNA-seq data). Due to standardisation (e.g. the MIAME standard for microarray data), it is relatively easy to exchange this type of data.

While repositories for microarray data have become very common, there are still very few databases providing spatial expression patterns. Two large-scale efforts are the Berkeley Drosophila Genome Project (BDGP) in situ database (www.fruitfly.org/cgi-bin/ex/insitu.pl) and the Edinburgh Mouse Atlas (genex.hgu.mrc.ac.uk). Both of these databases provide data for a large number of genes at a limited number of time points. The FlyEx database, on the other hand, provides data for a moderate number of genes involved in segment determination, but at a high spatial and temporal resolution (urchin.spbcas.ru/FlyEx).

These repositories provide an unprecedented wealth of data for modelling biological systems. But there are serious unresolved problems. Despite the rapid increase in available data, the measurements required for a specific modelling task are often missing or incomplete. Data heterogeneity poses further serious challenges for the model builder.

In this project, we aim at addressing problems such as these. We plan to provide **database infrastructure and standardise software tools** for integrating and combining different heterogeneous sources of data in a coherent and systematic way. We will establish workflows that allow the modeller to choose appropriate data from diverse repositories, such as pathway (e.g. KEGG, Reactome), protein-protein (MINT, IntAct, HPRD, STRING), transcription factor (TRANSFAC, Jasper), kinase-interactions (NetworKIN), and miRNA-target databases (miRBase) as well as expression data from sources such as ArrayExpress or GEO. These workflows should also allow the user to include predicted interactions from high-quality and experimentally validated computational software and text-mining, and to add annotation (e.g. on protein modifications, gene ontology etc) from genome databases (such as FlyBase, USCG and ENSEMBL). We will have direct access to these resources, many of which are located at EMBL-EBI. The collected data will be stored in the NetBase infrastructure developed at FTELE.IGM (unpublished).

On the foundation of the database structure described above, we will develop editors to create data files for modelling. We will combine standards for models (SBML; Hucka *et al.* 2003), model simulations (SBRML; Dada *et al.* 2010), and data (MIAME for gene expression; Brazma *et al.* 2001, MIAPE for proteomics, etc.).

We will then develop appropriate interfaces linking our databases to tools that will be used for analysis, visualisation and modelling. These include the tools to be developed during this project, as well as existing software such as CellNOpt (Saez-Rodriguez *et al.* 2009), a tool to construct logical models based on prior knowledge and high-throughput data and state-of-the-art visualisation tools (see Section 1.2.2 below).

1.2.2 Data Visualisation and Analysis

Model building requires both integration and abstraction, based on the complex datasets described in the previous section. We need to be able to identify the relevant components and interactions for our model. We need to see trends, and clusters of data points. We need to be able to filter out uninformative variables and identify anomalous or unusual data points (outliers). We need to pick out interesting features from the data at one glance. To enable such things, we need powerful tools for data visualisation and analysis, which go beyond what is already available in the scientific literature.

In this project we aim at advancing the state of the art developing tools that allow researchers to analyze the precise timing and localization of gene expression, compare spatio-temporal patterns across species, and visualize variability (e.g. within embryos and between embryos in developmental systems). We will build on expertise and existing tools within the consortium. For example, DataRail is a toolbox for managing, transforming, visualizing, and modelling data, in particular the multi-dimensional, high-throughput data encountered in systems biology (Saez-Rodriguez *et al.* 2008). We will extend DataRail (in a collaboration between EMBL, UShef and the Sorger Lab at Harvard) by including non-linear dimensionality reduction techniques such as the Gaussian Process Latent Variable Model (GPLVM, introduced by UShef; Lawrence 2005), which allow for visualisation of very high-dimensional datasets, which often contain non-linear low-dimensional structure. The GPLVM has now been extended to the analysis of time-series datasets and hierarchical models and has been successfully applied in a diverse set of domains, e.g. robotics, animation and tracking, but has not yet been widely applied in analysis of biological data or models. We expect that this and other recent developments from the field of machine learning will provide additional flexibility and power to DataRail.

While DataRail provides useful visualisation of high-throughput multivariate time-course data without spatial structure, we are also interested in developing tools for spatial time-course. We will therefore integrate existing methods developed in collaboration between partners CRG and UvA, which systematically analyze and compare spatial gene expression patterns.

The consortium also has unique expertise in low-level processing and data analysis of high-throughput data using probabilistic models. The puma (propagating uncertainty in microarray analysis) package allows robust model-based clustering, identification of significant multi-factorial trends and dimensionality reduction of high-dimensional data while properly accounting for the very noisy and heteroscedastic nature of gene expression data (Pearson et al. 2009). These tools are being extended to more diverse datasets such as RNA-seq and ChIP-seq and the current project

will provide a useful interface to these methods, which can be used to identify clusters and patterns in data without confounding by outliers and noisy variables.

1.2.3 The Model-Building Cycle

Although software tools exist for most of the individual steps in the model-building cycle (see Fig. 1), these tools are often neither straightforward to use, nor are they necessarily consistent and interoperable. Different algorithms (for parameter identifiability analysis, or parameter estimation, for example) are often implemented within distinct code frameworks. This makes comparison and application to different problems difficult. Moreover, many of these cutting-edge software tools use idiosyncratic, non-standard input-output data formats, which need to be tediously converted to combine them in an integrated workflow.

In this project, we will advance the state-of-the-art by developing novel methods, integrated software tools and workflows to support the full model-building cycle. This effort will specifically target new methods and tools for:

a. data integration, analysis and visualisation (see sections 1.2.1 and 1.2.2)
b. model building, with a special focus on multi-scale modelling,
c. robust parameter estimation via global, non-linear optimisation,
d. parameter identifiability analysis (theoretical, practical),
e. model validation,
f. model selection and model discrimination,
g. optimal experimental design,
h. design/comparison of measures e.g. for multi-objective optimisation,
i. uncertainty quantification.

Multi-Scale Modelling

From its origins in the 1990s (Broughton, 1999), multi-scale modelling and simulation has now turned into a focal point of attention across scientific and engineering disciplines. Many communities (ranging from physics and biology to medicine, finance, and engineering) are confronted with the problem of understanding multi-scale systems that are central to their field. The inherent complexity of biological systems is well recognised; they are multi-level systems that require a multi-disciplinary approach bridging a wide range of temporal and spatial scales (Sloot & Hoekstra 2010). Even biological phenomena in a single living cell span over a wide range of spatial and temporal scales and the number of molecular species involved can vary significantly.

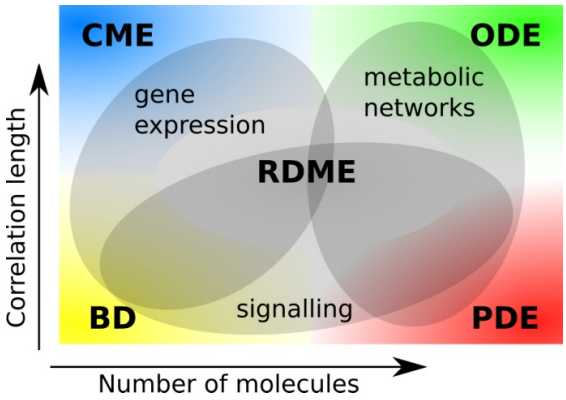

Figure 4: Regimes and models in biochemistry (Dobrzyński *et al.* 2007). Network models placed in correlation-length versus number-of molecules space. Abbreviations for (1) models with space: BD—Brownian dynamics, PDE—partial differential equation, RDME—reaction-diffusion master equation, and (2) models without spatial detail: CME—chemical master equation, ODE—ordinary differential equation. ODE and PDE are deterministic models; CME, RDME and BD are stochastic. The correlation length is a measure of the typical length scale at which a system retains its spatial homogeneity.

Current silicon cell platforms can often make reliable predictions for metabolic networks based on ordinary differential equations (ODEs; Fig. 4). For biochemical networks with membrane-bound molecules (e.g. signalling pathways), or in eukaryotic cells in general, methods based on partial differential equations (PDEs) are an appropriate approach. However, it is known that the process at the very origin of the whole cellular machinery, gene expression, gives rise to fluctuations in the concentration of the final protein products (Halford & Marko 2004, Becskei *et al.*, 2005). The discrete nature of matter under low-molecule-number conditions violates the continuum hypothesis used in ODEs and PDEs (Dobrzyński 2011). A model accounting for this is based on the chemical master equation (CME, van Kampen 1997), a deterministic linear ODE for the evolution of the probability density function for a Markov process. The CME approach remains valid as long as the

system is well-mixed. The question is whether this is a correct assumption when dealing with gene expression. Since there is a specific binding site, which needs to be found by a relatively small number of competing transcription factors, diffusion might limit the process thus giving rise to larger fluctuations (Metzler 2001). In order to resolve single diffusive encounters between bio-molecules a more detailed approach such as Brownian dynamics (BD, Allen & Tildesley, 2002) is needed. Unfortunately brute-force BD is too computationally expensive for large network simulations. More promising candidates for a versatile multi-scale framework are methods based on the reaction-diffusion master equation (RDME, Gardiner 1983)—an extension of CME for spatially distributed systems.

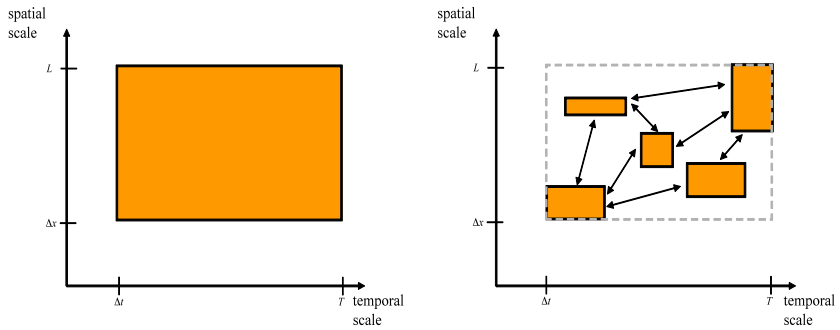

Figure 5: The Scale Separation Map and decomposition of a multiscale system: left, a multi-scale model spanning many temporal and spatial scales; right, the resulting decomposed model, consisting of 5 coupled single scale models.

In this project, we will systematically compare different frameworks to model biochemistry in terms of their ability to capture specific aspects of a system, and in terms of their interaction with algorithms for parameter estimation and model analysis.

For more complex systems we will use the multi-scale modelling methodology developed in the COAST project, coordinated by the UvA (Hoekstra 2010).Here, the building blocks of a multi-scale model are single scale models and their mutual multi-scale couplings. Many, if not all, multi-scale models lend themselves to such a partitioning strategy (Bassingthwaighte 2006, Sloot & Hoekstra 2010). The multi-scale model can be represented as a directed graph on a Scale Separation Map (SSM), which is a plot that has the relevant range of scales on its axes (usually space and time, but other quantities are possible). Single-scale models are positioned on the SSM according to their characteristic scales, and the coupling templates are represented as directed edges (Fig. 5).

Generic coupling strategies have been identified to interconnect several sub-models, each representing a different process at different spatio-temporal scales and corresponding to one component of the whole system. In addition to theoretical concepts, a coupling soft-ware environment (MUSCLE; www.berlios.de) has been developed and made available as open source, to build multi-scale applications. Within MUSCLE, both the kernels (i.e. the single-scale models) and the conduits (i.e. the multi-scale coupling) are software agents of the underlying multi-agent platform JADE (www.jade.tilab.com). The single-scale models do not need to be aware of each other, the information on the coupling and the global set-up are held by the framework. This allows the implementation of complex interfaces, where multi-scale couplings are performed by smart conduits. Furthermore, the structure of the coupling library allows complete independence from native codes. These can be replaced with a different source, provided the interface with the JAVA-wrapper agent remains the same.

The MUSCLE framework is currently being adapted by a number of EU projects, e.g. in the VPH domain (e.g. MeDDiCa) and in a project to realize Distributed Multi-scale Computing (MAPPER) applying MUSCLE again for VPH applications, but also in the field of computational biology, fusion, engineering, and nano-material science.

A challenging biomedical problem has been modelled with this approach, in-stent reste-nosis (Hoekstra 2010), which demonstrates the potential of this approach. We will use it in our project to integrate single-scale models on the molecular-, cellular- and tissue-level into multi-scale models and simulations of whole-cell, genomic regulatory networks in microorganisms, cell-cell signaling cascades, developmental gene regulatory networks involved in pattern formation, and integrated biotechnological production processes.

Model Selection/Model Discrimination

During the modelling process of biological (or other types of) phenomena it is not uncommon that several model frameworks can be proposed as descriptive for those phenomena (see above). A natural question that immediately arises is which model variant is best supported by the available

data. For deterministic models the current standard is that validation and model discrimination is based on statistical testing (for a review, see Cedersund & Roll 2009). A number of residual-based information criteria can be used, e.g. AIC and BIC (Akaike and Bayesian), the likelihood ratio test, and the F-test. Unfortunately, these criteria do not always result in a definitive answer, and their underlying assumptions may be invalid for complex, high-dimensional models. Bootstrapping—generating artificial data with the model to be tested against—can improve the reliability of the tests. Many of these statistical tests require that competing models are nested—i.e. have the same network structure—so they are especially useful for small model changes or to test various parameter vectors against each other.

If the model is of a probabilistic nature or the data are associated with a noise model then the support for the model given the data can be assessed through computation of Bayes factors. Bayesian inference has been shown to be a consistent framework for model comparison. A Bayes factor is the ratio of the probability for one model, M_1, given the data, D, to another model, M_2. The ratio of these probabilities $P(M_1|D) / P(M_2|D)$ gives an idea of which model is better supported by the data (Gelman *et al*. 1995). A further advantage of the Bayesian approach is that it gives a principled way in which further data sources can be integrated, for example, if the data from one experiment are denoted D_1 and an additional experiment is denoted D_2 the two data sets can be assimilated through Bayes' rule: $P(M_1|D_1, D_2) = P(D_2|M_1)\ P(M_1|D_1) / P(D_2)$. This allows for a cycle of model and experiment, where at each stage of the cycle new data from the experiment is assimilated with the existing knowledge. The principal difficulties associated with the approach are (1) encoding the modelling assumptions in a probabilistic manner, and (2) performing the necessary parameter integrals to compute the marginal likelihood of the model given the data. These two challenges interrelate: the more complex the probabilistic representation of the model, the more challenging the resulting integrals. However, the potential rewards are great and recent algorithmic advances mean that this Bayesian approach to model selection is now practically applicable to systems biology models (Vyshemirsky & Girolami 2008).The Bayesian approach deals naturally with parameter insensitivities (sloppiness in the parameters) through prior distributions. Parameters that are not identifiable simply retain the same distribution a posteriori (i.e. the posterior distribution) to their a priori specified distribution (the prior distribution). The presence of the non-identifiable para-meters is then easily checked through information theoretic measures of the dissimilarity between the prior and posterior distributions (such as the Kullback Leibler divergence, or information gain). The USheff group are world-leading in development of Bayesian models which integrate mechanistic assumptions (such as differential equations) but retain tractability such that potential network interactions can be validated through Bayes factors (Honkela *et al.* 2010).

Parameter Estimation by Global Non-Linear Optimisation

Given a specific modelling framework and a set of experimental data, we aim to calibrate the model. That is, we need to estimate parameter values, which cannot be measured directly, so as to fit the experimental results in the best possible way (Jaqaman and Danuser 2006). This is done by minimising a cost function, which measures the goodness of the fit (or alternative criteria, such as the robustness of the solution). Cost functions that have been shown to work well in practice include (i) the Bayesian estimator, (ii) the maximum likelihood estimator, and (iii) the (weighted) least squares estimator (Schittkowski 2002).

Estimating the parameters of non-linear dynamical models is difficult, since these models usually exhibit a large number of sub-optimal local minima (Schittkowski 2002). Traditional, local optimisation methods based on direct search or gradient descent are not suitable for such problems, since they tend to get stuck in these local minima. For this reason, there is no way of knowing if a bad fit is caused by a flaw or omission in the model formalism, or if it is simply a consequence of local convergence.

Therefore, we must resort to robust nonlinear optimisation techniques (Mendes & Kell 1998), which provide more guarantees of converging to the globally optimal solution (Moles *et al*. 2003). Examples of such techniques are simulated annealing, evolutionary algorithms, or scatter search (see, for example, Moles *et al.* 2003; Jaeger *et al.* 2004a,b; Fomekong-Nanfack *et al.* 2007; Rodriguez-Fernandez *et al.* 2006b). The importance of using global optimisation methods for

parameter estimation in systems biology has been increasingly recognized in recent years (Zwlolak *et al.* 2005; Tsai and Wang 2005).

Global optimisation methods can be roughly classified as deterministic, stochastic and hybrid strategies. Deterministic methods can guarantee—under some conditions and for certain problems—the location of the global optimum solution. Nevertheless, no determi-nistic algorithm can solve global optimisation problems of the class considered here with certainty in finite time. Stochastic methods are based on probabilistic algorithms, and they rely on statistical arguments to prove their convergence in a weak way. However, many stochastic methods can locate the vicinity of global solutions, but the associated computational cost is usually very large. In order to surmount this difficulty, hybrid methods and meta-heuristics have been recently presented (Rodriguez-Fernandez *et al*. 2006a) that speed up these methodologies while retaining their robustness.

The current challenge is how to perform parameter estimation in large-scale dynamic models. Although medium and large-scale dynamic models have been recently presented, these studies did not perform a proper full parameter estimation from experimental data (in most of them, subsets of kinetic data were chosen following ad hoc estimations, or were taken from the literature). Thus, there is a need to develop scalable parameter estimation methods, which are able to calibrate large-scale dynamic models of biological systems.

Measure Design: Cost Functions for Multi-Objective Optimisation

One major issue for parameter inference is that the observed dynamical behaviour of the system can often be explained by distinct regulatory mechanisms. This can be due to the optimisation problem being ill-posed or being insufficiently constrained by data (Ashyraliyev *et al.* 2009a). Alternatively, parameters can be difficult to determine due to correlations between them (Gutenkunst *et al.* 2007; Ashyraliyev *et al.* 2008; 2009b). Model discrimination based on additional experimental evidence is required to decide, which of the alternative mechanisms is applicable to the real biological system (see also above). This is often time-consuming and technically challenging. Therefore, it is essential to increase the reliability of the model results for experimental design and to decrease the number of alternative predictions that need to be tested experimentally.

One way of achieving this is to change the metric that measures the distance between the experimental data and the model results. Usually, the accuracy with which a model reproduces observed expression patterns is measured by a cost function based on the sum of squared differences between model and data (single-objective optimisation). The first option is to change the metric based on uncertainty quantification (UQ). Following Oberkampf & Barone (2006) a metric should take into account (i) the simulation error, (ii) the predictive accuracy of the model (obtained by UQ), (iii) the number of experimental measurements, (iv) the experimental measurement errors, and (v) the error resulting from data post-processing. Contrary to these authors recommendation we do not exclude “adequacy indications” in the metric, like e.g. robustness, but we stress the importance to redesign model validation, discrimination and experimental-design procedures based on the new metric. This is a challenging task, but required for the reliability of the predictions.

Another way to distinguish between alternative mechanisms is to include additional objectives in the metric (multi-objective optimisation) (Handl *et al.* 2007). For example, we can take advantage of the fact that biological regulatory processes must proceed reliably in the presence of molecular fluctuations, genetic variability and environmental perturbations. In other words, realistic biological processes are robust, and robustness should be considered when fitting models to data. Preliminary efforts have been made to apply multi-objective optimisation to reverse-engineering gene networks (van Someren *et al.* 2003; Esmaeili *et al.* 2009; Guo *et al.* 2009). In this project, we will extend these efforts in a systematic way.

Parameter Identifiability Analysis

Before performing the optimisation to infer the model-parameters from the experimental data one would like to know *if* the parameters can be determined at all, assuming that for all observables continuous and error-free data are available. This is the subject of *a priori* or *structural identifiability* analysis. For linear models the *Laplace transform* approach can be used (Godfrey & Fitch 1984), for nonlinear models the oldest method is the *Taylor* or *power series expansion* (Pohjanpalo 1978).

Another classical method is the *similarity transformation* (Vajda *et al.* 1989). Recently methods have been developed that use differential algebraic techniques (Audoly *et al.* 2001) which are implemented in the symbolic language REDUCE in a publicly available software tool DAISY (Bellu *et al.* 2007). Although this brought *a priori analysis* within reach of the biologists (see e.g., Roper *et al.* 2010), for realistic large-scale models it is still very difficult to obtain results. Therefore one often has to rely on *a posteriori identifiability* analysis once the parameters have been estimated. This analysis studies the influence of accuracy and sufficiency of the experimental data on the uncertainty in the model parameters. The most applied method to study this uncertainty in the parameters is to compute the Fisher Information Matrix (FIM, Hydalgo & Ayesa 2001) evaluated for the given data points and the parameter vector obtained by the data fit. The FIM describes an ellipsoidal confidence region, from which confidence intervals and correlations can be computed.

This analysis is easy and cheap to perform but it is also linear and local with respect to the parameters. A nonlinear analysis can be performed by Monte Carlo sampling of the parameter space around the parameter vector. Hengl *et al.* (2007) propose another interesting nonlinear analysis: repeated fitting for different initial guesses of the parameter vector. The resulting parameter vector matrix is then analyzed with Alternating Conditional Expectation (Breiman & Friedman 1985) resulting in optimal transformations for the parameters to come to an identifiable model.

Finally, as stated above, Bayesian parameter estimation provides a natural framework for assessing parameter identifiability. The prior and posterior parameter distributions capture the uncertainty in parameter estimates before and after observing some data (simulated from the model or experimental). Parameters, which are difficult to identify are associated with a small difference between the prior and posterior, usually quantified by the Kullback-Leibler divergence or some other convenient divergence measure. An advantage of the Bayesian approach is that we can retain this uncertainty information for poorly defined parameters and model predictions are made by integrating over the distribution of parameters, rather than making an arbitrary choice between unlikely specific parameter sets. Parameter identifiability is intimately related to optimal experimental design since we would like to select experiments that are informative with respect to important model parameters. Again, this can be achieved in a Bayesian context by selecting experiments that are most likely (given the current posterior distribution over models) to improve our knowledge about parameters of interest. In this project we will extend Bayesian parameter identifiability analysis to differential equation models over graphs. This is a challenging problem since the model likelihood will be expensive to compute, requiring numerical integration, and we will therefore investigate speed-ups to avoid excessive simulations.

Model validation

Model validation checks whether the model agrees with the biological data/evidence. Most often this is done in a qualitative way based on the inferred model, e.g. by graphical inspection of the model results versus the experimental data or by looking at the residual of the objective function. A rigorous model validation however requires either independent validation data or *cross-validation* (Geisser 1993). The reason is that the inferred model will in general better fit the “training” data than any other independent sample of the data (*over-fitting*). Cross-validation predicts the model-fit to a hypothetical validation set when an explicit validation set is not available. A common type of cross-validation is resampling (repeated splitting of the data in training/validation data).

A good measure of the distance between model and data will enhance the reliability and the functionality of the model validation (Oberkampf & Barone 2006).

A general validation procedure in a Bayesian setting is proposed in Babuška *et al.* (2008). Here the validation data is used to produce a Bayesian update of the model and the distance between the updated model and the original one determines whether the model is acceptable.

Optimal Experimental Design

Performing experiments to obtain a rich enough set of experimental data is costly and time-consuming. For this reason, Optimal Experimental Design (OED; Kreutz & Timmer 2009) is a critical step in the systems biology model building cycle (Fig. 1). The purpose of OED is to devise experiments in such a way that model parameters can be estimated from the resulting experimental data with the best possible statistical quality, which is usually a measure of the

accuracy and/or de-correlation of the estimated parameters (Kutalik *et al.* 2004; Gadkar *et al.* 2005; Kremling *et al.* 2004; Feng *et al.* 2006; Casey *et al.* 2007; Banga & Balsa-Canto, 2008). In other words, based on candidate model frameworks, we seek to design the best possible experiments in order to facilitate system identification. To achieve this, OED relies on statistical analysis and optimisation techniques. While OED applied to linear steady-state models is a well-established subject, OED of non-linear dynamic models is more challenging and no satisfactory methods are available at his point.

Several slightly different criteria for OED—denominated by an alphabetic nomenclature (Kiefer 1959)—are defined for this purpose. All of these are based on the Fisher information matrix. After selection of a suitable criterion, different approaches can be used for obtaining an optimal experiment. One approach—followed by Melas (2006)—tries to transform the problem into a Chebyshev system. From this system, a Chebyshev polynomial is constructed, which is used to base the experiment on. Another approach—used by Asprey & Macchietto (2002), Balsa-Canto *et al.* (2008) and Bauer *et al.* (2000)—converts the problem into a (semi-infinite) optimisation control problem.

In this project we will formulate the general problem as a mixed-integer dynamic optimisation (MIDO) problem, and will develop algorithms for its numerical solution. These algorithms for MIDO can be obtained using direct methods, which transform the original problem into a mixed-integer-nonlinear programming (MINLP) problem via parame-trisations of the controls and/or states. However, because of the frequent non-smoothness of the cost functions, the use of gradient-based methods to solve this NLP might lead to local solutions. As for parameter estimation (see above) there is a need of global optimisation methods to ensure proper solutions. Stochastic methods for global optimisation are the most robust methods for this class of problems (Banga & Balsa-Canto, 2008). However, the challenge remains to apply these methods to realistic, large-scale kinetic models of biological systems.

Another approach we will consider is Bayesian optimal experimental design. Above we described the Bayesian perspective on parameter identifiability, which is based on assessing the difference between prior and posterior distributions after observing some data. We can also use Bayesian methods to investigate the expected change in the posterior distribution given an experiment or a sequence of experiments by averaging over the experimental outcomes given current beliefs (captured by the current posterior distribution). This allows us to seek experiments producing the largest expected information gain. This relates also to the problem of choosing an appropriate measure (which parameters or model outputs to focus on) and it would be interesting to explore the relationship between Bayesian methods and multi-objective methods by considering Bayesian inference applied over a range of cost functions.

Uncertainty Quantification

Uncertainty quantification (Karniadakis & Glimm 2006, Ghanem & Wojtkiewicz 2004) studies the propagation of numerical errors and model errors caused by e.g. limited data, sloppy parameters, inaccurate input values, etc. The classical statistical approach for UQ is *Monte Carlo*. The parameter space is sampled and model simulations, with each parameter drawn from its uncertainty distribution, produce an ensemble of random results. This results in a probality density function for the outcome. The convergence, however, of this process is very slow. Accelerating techniques are a.o. Markov Chain Monte Carlo (Smith 1984), and Latin hypercube sampling (McKay *et al.* 1979). A more economical approach is the *sensitivity method* that is based on moments of samples, but this is less robust.

A non-statistical method, the *Polynomial Chaos (PC) expansion* (Ghanem & Spanos 1991) has been used frequently in the last years. It is based on an hierarchical representation of the stochastic process (like spectral expansions). The PC method or its variants have been applied to a number of applications, a.o. to stiff systems (Cheng & Sandu 2009).

In this project both the statistical and the non-statistical approach will be used. The latter is more suitable for systems with a small number of uncertainties with a large variance in the value; the former is conceptual simple, but requires a HPC implementation.

1.2.4. Application: Large-scale Models of Microorgansims and Eukaryotic Cell Lines

In this project, we aim to develop mathematical and computational strategies to create large-scale models using data from multiple sources. This includes metabolic networks as well as gene regulatory networks. The final aim of constructing such detailed models is to exploit them in biotechnological applications, typically making use of computer aided metabolic engineering procedures.

The structure of metabolic networks is approachable by a reconstruction approach using data from genome annotation, metabolic databases and chemical databases such as ChEBI and KEGG (Palsson & Thiele, 2010). In addition to the structure of the network, we then proceed to set generic rate laws to represent the kinetics of all algorithms and finally fill in details of the precise mechanisms of those reactions that are known in detail. This strategy leads to a kinetic model that is as accurate as current knowledge allows, which can be explored using various modelling analyses coming from the methods developed in this project. The application of these models to biotechnology has a wide application range, for example for metabolic engineering, where existing metabolic pathways are altered to increase yield and/or flux of compounds of commercial interest. Another area where these models are useful in biotechnology is in optimization of strains and culture conditions for improved production of biopharmaceuticals. In both of these cases, and others, kinetic models allow us to identify multiple points in the network which can be modulated for optimal production. Stochiometric models, such as flux balance analysis, even though very useful, provide only a limited level of prediction with little or no extrapolation power. Metabolic kinetic models, which are obtained by adding kinetic rate laws with appropriate parameter values, are much more informative because they provide extrapolation power, however they are only appropriate while genetic regulation is not relevant. To be fully predictive one needs to extend the metabolic kinetic model to include gene regulation; the combination of kinetic models and gene regulatory models is thus of great importance to biotechnology.

We have experience of developing metabolic reconstructions and further develop them to large kinetic models, having applied this process to *Saccharomyces cerevisiae* (Smallbone *et al.* 2007; Herrgard *et al.* 2008; Dobson *et al.* 2010; Smallbone *et al.* 2010). Here we will continue developing the approach while applying it to other organisms of biotechnological interest: *S. cervisiae, E. coli* and Chinese Hamster Ovary (CHO) cells. Development of these new large-scale models is not without challenges: while there are reconstructions of metabolism of *E. coli* already, there is no large-scale metabolic kinetic model for this organism, let alone a combined metabolic and gene regulatory network model. The same applies to CHO cells, with the added challenge of being higher eukaryote cells.

An important area of research in constructing these large kinetic models, is the choice of kinetic rate laws to use for each reaction mechanism. We have previously used the lin-log kinetic type but have continued studying several other options, such as convenience kinetics and other generic rate laws formulated with similarity to mechanistic enzyme kinetics rate laws (Liebermeister *et al* 2010). In any case, the model will contain a large number of parameters that must be estimated. We start by collecting information about the thermodynamics of reactions with estimates of equilibrium constants. This is followed by a global fit to data from reaction fluxes and metabolite concentrations, obtained from flux balance analysis and metabolomics studies. This parameter estimation exercise is carried out using our methods including stochastic and hybrid global optimisation algorithms (Mendes & Kell 1998, Rodriguez-Fernandez et al. 2006a). Finally the results of parameter estimation are followed by parameter sensitivity and identifiability analysis to uncover which ones may need more accurate estimates, and those which the model is robust against. In particular, it is important to carry out global sensitivity analysis, for which we also use optimisation methods (Sahle *et al.* 2008).

We have also considerable experience in reverse-engineering gene regulatory networks (Bansal et al. 2007). As opposed to earlier studies, our goal is to learn a dynamic model of the network, rather than a static network map, by analyzing massive experimental datasets, including all the available gene expression data and taking into account prior knowledge. For these reasons, it will be necessary to also consider a probabilistic framework of gene interaction. In such a framework, the model M is learned from data D, by maximising a probability function, which can be converted to an equivalent problem using Bayes' rule which naturally includes prior knowledge. Also once we

have learned the model M, for a new dataset D_1 we can ask what is the probability that D_1 has been generated by model M. The most general method in this category consists of dynamic Bayesian networks.

Bayesian networks, however, do not scale well with increasing biological system size, due to the heuristic step required when identifying the correct model M. In this step the Bayesian network approach needs to try different topologies of the network. Although the method does not need to search them exhaustively, it has to search a large enough space to be sure that a good solution is found. Due to the sheer size of the network (>20,000 genes), even searching a small space is challenging with current computational power.

Another approach, which can deal with such complexity, consists of association networks based on mutual information (MI) (e.g. Margolin *et al.* 2006). In this case, model space is restricted to pair-wise interactions. One limitation of this simplification is the loss in the ability to identify direct interactions, as compared to indirect interactions. This approach lacks the ability to include prior knowledge, as well as the ability to interpret new data.

In order to overcome the limitations of these methodologies, we need to develop a novel method that satisfies all of our required features: scalability, inclusion of prior information, and particularly the requirement of being able to interpret new data. In order to obtain a predictive dynamic model able to satisfy the required features, we will explore a Bayesian approach, in which we will learn a probabilistic model for each pair-wise (and possibly three-way) interaction across all the genes using all the information available in NetBase (see Section 1.2.1), thus overcoming the problem of learning massive networks using classical Bayesian approaches. Each pair-wise interaction will be modelled as a continuous, or discrete, probability distribution, whose unknown parameters will be learned from the expression data and from prior knowledge. Prior knowledge in NetBase will be captured by setting a prior distribution on the parameters to be learned. We will need to investigate the most appropriate functional form of the prior distribution, depending on the kind of prior knowledge. For each interaction between two genes, we will learn a general hierarchical Bayesian model. We plan to use a Monte Carlo Markov-Chain approach to find the posterior probability of the parameter(s) of the probability distribution, from the observed expression data, and from prior knowledge.

This new methodology will be tested by application to the creation of a dynamic gene regulatory network model of E. coli. Subsequentely this gene regulatory network will be integrated into the kinetic model of metabolism resulting in a comprehensive predictive model of E. coli, which will be an invaluable resource for biotechnology. Such a model allows to predict the effects that are not just limited to metabolic regulation, but also to to responses that include altered gene expression. While not being a fully mechanistic model (i.e. may not include significant aspects of the mechanistic details of the underlying molecular interactions), such a model is much more than a phenomenological model, and can be seen as a stepping stone towards a global (systems) understading of the biochemistry and genetics of one of the most important host cells for biotechnology.

1.2.5 Application: Signalling and Regulatory Networks in Eukaryotic Cells

Modelling signalling and regulatory networks is very challenging, due to the large number of molecules involved, the highly non-linear and dynamic behaviour often observed, and the difficulty to obtain quantitative measurements. Typically, models cover only one or two pathways, which are modelled using differential equations to describe the underlying biochemistry (Chen 2009). Recently, rule-based approaches have been developed as a means to deal with the inherent combinatorial complexity (Hlavacek *et al.* 2006). Novel experimental techniques such as protein arrays, bead-based systems (e.g Luminex) and mass spectrometry provide large amount of data about signalling processes. Therefore, it has become possible to probe larger signalling networks under multiple conditions. Additionally, prior knowledge information is becoming increasingly available in an integrated manner, thanks to efforts to unify pathways description (Biopax; www.biopax.org) and to create a common portal to access different databases (PathwayCommons; www.pathwaycommons.org/pc).

With this large amount of data and network information, it is in principle possible to generate models of large signalling networks. However, to identify the exact network structure and the

kinetic parameters poses an enormous optimisation problem. Due to this, efforts so far have attempted to model these networks using simple formalisms such as Boolean or fuzzy logic (Morris *et al.* 2010). Such methods, however, only provide an extremely simplified description of the underlying biochemistry.

The EMBL group has extensive expertise on modelling signalling networks, and the FTELE.IGM group on modelling regulatory networks. The EMBL group has developed a framework to model large signalling networks using discrete logic, embedded in the tool CellNetOptimizer (CellNOpt; Saez-Rodriguez *et al.* 2009), and is currently extending the approach for continuous, dynamical systems. The group also has experience in modelling signalling networks with biochemical formalisms. These different formalisms will provide suitable benchmarks for the methods for model selection and parameter estimation. We will use the NetBase platform to infer prior knowledge networks (in combination with available resources such as PathwayCommons). The resulting networks will be trained with CellNOpt using high-throughput proteomics data collected by collaborators of the EMBL group.

We will focus on relevant signalling and regulatory networks, using data on cell lines that are commonly used in biotechnology as models to develop novel drug therapies, toxicity studies, etc. Furthermore, we will link these models to models of metabolism in cell lines, in particular, of the CHO cell line. Insilico Biotechnolgoy has developed a comprehensive kinetic model of metabolism in these cells, and we will collaborate on connecting these models. The models developed in this section can be used as in silico tools for the optimisation of biotechnological production processes (see Section 1.2.7 below).

1.2.6 Application: Spatial Models of Gene Regulatory Networks in Development

Today, a large majority of industrial biotechnological production processes are carried out in unicellular microbial systems. On the other hand, genetically engineered plant systems have a huge potential for biotechnological applications, such as the production of biofuels or bio-degradable plastics. Similarly, genetic engineering in farm animals is of increasing economic importance. However, the genetic manipulation of complex, multi-cellular orga-nisms is still in its infancy, since our current methods of intervention are crude, and we lack a rigorous, quantitative understanding of the complex regulatory networks involved in animal or plant growth and physiology. Such understanding would allow more fine-tuned, well-adjusted, and more effective genetic engineering (and synthetic biology) in multi-cellular systems. In particular, it would enable us to express relevant transgenic factors at exactly those points in space and time at which they are required for a specific application.

Gene networks acting during development in multi-cellular organisms pose special challenges for reverse-engineering and modelling. In contrast to the large-scale microbial and signalling networks considered so far, developmental and physiological processes usually only involve a moderate number of genes, but exhibit highly intricate spatial and temporal regulatory dynamics. The related optimisation problems are extremely complex, and provide a tough challenge for our global optimisation algorithms and related methods.

We will consider two particular cases of developmental systems as benchmark problems. Both are involved in pattern formation in early animal development. These systems are representative of many regulatory networks in biology: the insights gained from such an analysis, and the technical challenges posed by these systems, can be easily generalised.

Our first choice of model system is early development of the fruit fly *Drosophila melano-gaster.* The gene networks underlying pattern formation in *Drosophila* during the first few ours of embryogenesis are probably the best-studied developmental gene regulatory networks available at the moment. This allows us to rigorously compare modelling results with the high-resolution quantitative datasets of spatial gene expression patterns available for this system, which have already been used to infer the regulatory dynamics of pattern formation using a reverse-engineering approach (Reinitz & Sharp, 1995; Jaeger et al., 2004a,b; Perkins et al., 2006; Manu et al., 2009a,b; Ashyraliyev et al. 2009b).

The CRG group has been extending this approach to a comparative analysis of network evolution. In particular, we have created quantitative datasets of spatial gene expression patterns for three species of dipterans (flies, midges and mosquitoes). These datasets form a unique platform for

reverse-engineering and comparing pattern-forming networks between species. We will obtain models of segmentation gene expression based on different modelling formalisms using the modelling cycle (Fig. 1). Analysis of the resulting models will be used to identify similarities and differences in the dynamical behaviour of the system, which can explain commonalities and divergence of gene expression between species. Understanding how a gene regulatory network can be altered during evolution, will aid our understanding of how to engineer complex spatial networks in the future.

We will also consider the starlet sea anemone (*Nematostella vectensis*) an emerging model system for the experimental study of development and evolution (Finnerty *et al.* 2004; Kusserow *et al.* 2005). The UvA group has developed quantification methods to measure spatial profiles of gene expression during early developmental stages of *Nematostella* and has been involved in developing models for pattern formation and morphogenesis in this species (Tamulonis *et al.* 2011). Based on these pioneering efforts, we plan to use early *Nematostella* development as a test case for the inference and modelling of pattern-forming regulatory networks.

Nematostella is more representative of many developmental processes in other animals than the fly systems described above: First, our knowledge, and hence the datasets used for model inference, are still much more preliminary and incomplete than in the case of *Drosophila.* Therefore, this system will test the ability of our methods to cope with noisy and uneven datasets. And second, early development of *Nematostella* occurs in cellularised tissues, involving cell movements and signalling between different tissues. It is therefore more representative and less derived than early dipteran development.

1.2.7 Application: Production Processes in Industrial Biotechnology

The application of data-driven mathematical models in industry for the improvement of biotechnology production processes has only just begun, and the regular use of modelling and optimisation software in the private biotechnology sector needs to be promoted. The use of such methods is often hampered by the absence of user-friendly, flexible and reliable software. Existing code often needs significant expert knowledge (both computational and scientific). In other words, end-users without extensive expertise in how to handle and compile code, and in modelling and optimisation, are excluded from the use of advanced algorithms and models, or simply need too much time to use and understand their functionality. Hence, there is a strong demand for user-friendly software solutions implementing the iterative modelling cycle described above that can be used by non-experts and guide the design of efficient production processes.

One of our main aims in this project is the development of user-friendly software to support the model-building cycle. This will be achieved by a close interaction of academic partners (who are developing the algorithms), end-users (Insilico Biotechnolgy, INSIL and Fluxome SA, FS; who will be applying the software in a commercial biotechnological setting) and Complex Systems Modelling (CSM; who will be in charge of writing our integrated code framework). Users with different levels of expertise (experimental biologists, engineers, and bioinformaticians) will be employed to test our emerging software framework and to provide feedback on user-friendliness and functionality to the developers. We aim at establishing an efficient process of knowledge transfer for the academic partners to the biotechnology SMEs.

The functionality of the software will be tested by application of models to various organisms used for the production of compounds of high industrial interest, including nutraceutical, ingredients, biopharmaceuticals, and fine chemicals. In particular, we will focus a) on the simulation of the production of nutraceutical ingredients such as resveratrol and polyunsaturated fatty acids using dynamic models of *S. cerevisiae*, b) on the simulation of therapeutic antibody production using dynamic models of Chinese Hamster Ovary (CHO) cells, and c) on the simulation of amino acid production using dynamic models of *Escherichia coli*. Simulation results are likely to lead to the identification of novel metabolic engineering and synthetic biology targets that can improve the production efficiency of the compounds under investigation.

Our second main aim in this part of the project is to go beyond the use of steady-state genome-scale models that have recently been shown to give promising results for metabolic engineering i.e. in lycopene and ethanol production (Alper *et al.* 2005; Bro *et al.* 2006). These models do not yet include any regulatory or dynamic information, and simulation capabilities can become rapidly

limiting. It is expected that our efforts will lead to dynamic models that will yield superior and more accurate simulation results. This would enable and boost the increasingly widespread use of such models in the design of biotechnological production processes.

Such dynamic simulations capitalize on network models combining the interaction of metabolism, gene regulation and/or signalling processes. For the production of therapeutic antibodies using CHO cell cultures, for example, large-scale dynamic models will pave the way for predicting the impact of relevant process variables like pH and/or media composition on cell growth and productivity or regarding clinically important aspects of product quality, such as glycosylation patterns. Such predictions are notoriously difficult or unreliable using today's methods.

1.3 S/T Methodology and Associated Work Plan

i) Describe the Overall Strategy of the Work Plan

We have subdivided our work plan into work packages (WP) as follows:

The three vertical objectives of the project (see section 1.1.7 and Fig. 3) are represented by WP1–3. WP1 and 2 deal with tools/algorithms for data analysis and visualisation respectively. They are prerequisites for the application of the modelling cycle tools to be developed in WP3, which depend on the availability of data repositories to enable data-driven modelling, model validation, parameter estimation, uncertainty quantification and optimal experimental design. WP3 constitutes the central effort of the project, involving all the academic partners, and one of the SMEs. WP1–3 will provide tools and methods, which will be tested using different biological and biotechnological applications in WP4–7.

The horizontal objective is subdivided into four areas of applications (Fig. 3), which are implemented in a separate work package each: WP4 deals with large-scale metabolic and gene regulatory network models for micro-organisms and cell lines. It is a prerequisite for WP7 (biotechnological applications), for which such models will be needed. WP5 deals with modelling cell-cell signalling cascades, and WP6 with developmental gene regulatory networks. Both of these applications introduce aspects of models (such as spatially distributed systems, and extremely heterogeneous sources of data), which are representative for many biotechnological, and in particular, many future synthetic biology applications. The most important work package in this objective is WP7. It implements the application of the methods developed in WP1–3 to biotechnological production processes and will provide a close collaboration between tool developers (academic partners) and users of these tools (SMEs).

We dedicate separate work packages to complementary (but crucial) activities associated with the primary research effort of the consortium. The main aim of our project is to develop and implement novel methods for modelling and optimisation. WP8 will be concerned with the dissemination of these methods and potential exploitation of project results. It is led by one of the SMEs (CSM) who will distribute our software and will provide maintenance and support to users and customers. WP8 covers not only the legal infrastructure for CSM to implement their code, but also includes activities such as presentations at scientific meetings and specialised biotechnology and bio-IT trade fairs.

Another work package is dedicated to training within and outside our consortium. WP9 deals with the organization of workshops (one early, internal one; and a later, publicly accessible one) and will be responsible for organising exchanges of researchers between the partners of the consortium. This will provide unique training opportunities to the young researchers involved, and will also facilitate the communication and information exchange between partners.

Finally, we dedicate a separate work package (WP10) to the management of the project. It will ensure that legal requirements are met, that reports are delivered in a timely fashion, and that an efficient flow of information is established between the partners and the work packages. In addition, WP10 will be in charge of organising regular meetings of the steering committee, the general assembly, and the scientific workshops organised by the consortium. This work package will be implemented by the International Collaborations Office (ICO) at the CRG, which is dedicated to the management of European projects.

ii) Show the Timing of the Different WPs and their Components (Gantt Chart)

	Work Package Title		Deliverable Title	Year 1												Year 2												Year 3											
WP1	Database Integration & Exploitation	1.1	Database Infrastructure																																				
		1.2	Database/Tools Interface																																				
		1.3	Integration Workflows																																				
		1.4	Data Integration Tools																																				
		1.5	Model Data File Editor																																				
WP2	Visualisation Tools for Data & Model Building	2.1	GPLVM Software																																				
		2.2	DataRail Visualisation Tools																																				
		2.3	Spatial Visualisation Tools																																				
WP3	Integrated Software Tools for the Modelling Cycle	3.1	Bayesian Inference Tools																																				
		3.2	Parameter Estimation Tools																																				
		3.3	Multi-objective Optimisation Tools																																				
		3.4	Integrated Suite of Tools																																				
WP4	Application: Large-scale Models of Microorganisms	4.1	Reconstruction of *E. coli* metabolism																																				
		4.2	Genome-wide Kinetic Model of *S. cervisiae*																																				
		4.3	Reconstruction of CHO Cell Metabolism																																				

Each year is subdivided into 12 periods of one month.

ii) Show the Timing of the Different WPs and their Components (Gantt Chart) (Continued)

	Work Package Title		Deliverable Title	Year 1											Year 2												Year 3											
WP4 (ctnd.)	Application: Large-scale Models of Microorganisms (ctnd.)	4.4	Genome-wide Kinetic Model of *E. coli*																																		
		4.5	Gene Regulatory Network of *E. coli*																																		
		4.6	Combined Metabolic/Regulatory Model of *E. coli*																																		
WP5	Application: Signalling & Regulatory Networks in Cells	5.1	Algorithms for Integration of Signalling Data																																		
		5.2	Reconstruction of CHO Signalling Networks																																		
		5.3	Kinetic Models of CHO Signalling Networks																																		
		5.4	Integrated Signalling/Metabolic Models (CHO)																																		
WP6	Application: Developmental Gene Regulatory Networks in Animals	6.1	Datasets for Spatial Gene Expression																																		
		6.2	Animal Regul. Network Models																																		
WP7	Application: Biotechnological Production Processes	7.1	Specifications for Software Functionality & GUI																																		
		7.2	Prototype Software for Testing																																		
		7.3	Models: Biotechnological Production Processes																																		
		7.4	Comparative Analysis of Producer Strains																																		
		7.5	Target Identification for Process Optimisation																																		

Each year is subdivided into 12 periods of one month.

ii) Show the Timing of the Different WPs and their Components (Gantt Chart) (Continued)

	Work Package Title		Deliverable Title	Year 1											Year 2												Year 3											
WP8	Dissemination & Technology Transfer	8.1	Project Website																																		
		8.2	Software Development/Testing Architecture																																		
		8.3	Integrate Software Suite																																		
		8.4	Talks/Demo Stalls at Meetings																																		
		8.5	Manuscripts on Software Suite/Tools																																		
WP9	Training	9.1	Internal Workshop at the CRG																																		
		9.2	External Workshop at the EBI/EMBL																																		
		9.3	Researcher Exchange Visits between Partners																																		
WP10	Project Management	10.1	Consortium Agreement																																		
		10.2	Quality Assurance Plan																																		
		10.3	Kick-off Meeting																																		
		10.4	1st Annual Meeting																																		
		10.5	1st Annual Activity & Management Report																																		
		10.6	2nd Annual Meeting																																		
		10.7	2nd Annual Activity & Management Report																																		
		10.8	Final Meeting																																		
		10.9	Final Activity & Management Reports																																		

Each year is subdivided into 12 periods of one month.

iii) Provide a Detailed Work Description Broken Down into Work Packages:

Table 1.3a: Work Package List

Work Package No.	Work Package Title	Type of Activity	Lead Participant No.	Lead Participant Short Name	Person-Months	Start Month	End Month
WP1	Database Integration & Exploitation	RTD	3	FTELE.IGM	41	1	18
WP2	Visualisation Tools for Data & Model Building	RTD	8	USheff	20.2	1	12
WP3	Integrated Software Tools for the Modelling Cycle	RTD	2	CSIC	139	1	36
WP4	Application: Large-scale Models of Microorganisms	RTD	7	UNIMAN	47	1	36
WP5	Application: Signalling & Regulatory Networks in Cells	RTD	6	EMBL	46	12	36
WP6	Application: Developmental Gene Regulatory Networks in Animals	RTD	4	UvA	50.6	12	36
WP7	Application: Biotechnological Production Processes	RTD	12	FS	55	1	36
WP8	Dissemination & Exploitation	DEM	9	CSM	23	1	36
WP9	Training	OTHER	1	CRG	6	1	36
WP10	Project Management	MGT	1	CRG	10	1	36
				TOTAL	437.8		

RTD: Research and Technology Development; DEM: Demonstration; MGT: Management.

Table 1.3b: Deliverables List

Del. No.	Deliverable Name	WP No.	Nature	Dissemi-nation Level	Delivery Date
1.1	Database Infrastructure	WP1	P	PU	6
1.2	Database/Tools Interface	WP1	P	PU	6
1.3	Integration Workflows	WP1	P	PU	12
1.4	Data Integration Tools	WP1	P	PU	12
1.5	Model Data File Editor	WP1	P	PU	18
2.1	GPLVM Software	WP2	P	PU	12
2.2	DataRail Visualisation Tools	WP2	P	PU	12
2.3	Spatial Visualisation Tools	WP2	P	PU	12
3.1	Bayesian Inference Tools	WP3	P	PU	18
3.2	Parameter Estimation Tools	WP3	R/P	PU	18
3.3	Multi-objective Optimisation Tools	WP3	P	PU	24
3.4	Integrated Suite of Tools	WP3	R/P	PU	36
4.1	Reconstruction of *E. coli* metabolism	WP4	R/P	PU	12
4.2	Genome-wide Kinetic Model of *S. cervisiae*	WP4	R/P	PU	12
4.3	Reconstruction of CHO Cell Metabolism	WP4	R/P	PU	12
4.4	Genome-wide Kinetic Model of *E. coli*	WP4	R/P	PU	18
4.5	Gene Regulatory Network of *E. coli*	WP4	R/P	PU	24
4.6	Combined Metabolic/Regulatory Model of *E. coli*	WP4	R/P	PU	36
5.1	Algorithms for Integration of Signalling Data	WP5	P	PU	18
5.2	Reconstruction of CHO Signalling Networks	WP5	R/P	PU	24
5.3	Kinetic Models of CHO Signalling Networks	WP5	R/P	PU	30
5.4	Integrated Signalling/Metabolic Models (CHO)	WP5	R/P	PU	36
6.1	Datasets for Spatial Gene Expression	WP6	P	PU	18
6.2	Animal Regulatory Network Models	WP6	R/P	PU	36
7.1	Specifications for Software Functionality & GUI	WP7	R	PP	3
7.2	Prototype Software for Testing	WP7	P	PP	6
7.3	Models: Biotechnological Production Processes	WP7	R/P	PP	24
7.4	Comparative Analysis of Producer Strains	WP7	R	PP	30
7.5	Target Identification for Process Optimisation	WP7	R	PP	36
8.1	Project Website	WP8	D	PU	6
8.2	Software Development/Testing Architecture	WP8	P	PP	6
8.3	Integrate Software Suite	WP8	P	PU	36
8.4	Talks/Demo Stalls at Meetings	WP8	R/D	PU	36
8.5	Manuscripts on Software Suite/Tools	WP8	R	PU	36

P: Prototype; R: Report; D: Demonstrator; O: Other; PU: Public; PP: Shared only Among Participants

Table 1.3b: Deliverables List (contd.)

Del. No.	Deliverable Name	WP No.	Nature	Dissemi-nation Level	Delivery Date
9.1	Internal Workshop at the CRG	WP9	D	PP	12
9.2	External Workshop at the EBI/EMBL	WP9	D	PU	36
9.3	Researcher Exchange Visits Between Partners	WP9	O	PP	36
10.1	Consortium Agreement	WP10	O	PP	3
10.2	Quality Assurance Plan	WP10	O	PP	3
10.3	Kick-off Meeting	WP10	O	PP	3
10.4	1st Annual Meeting	WP10	O	PP	12
10.5	1st Annual Activity & Management Report	WP10	R	PP	12
10.6	2nd Annual Meeting	WP10	O	PP	24
10.7	2nd Annual Activity & Management Report	WP10	R	PP	24
10.8	Final Meeting	WP10	O	PP	36
10.9	Final Activity & Management Reports	WP10	R	PP	36

P: Prototype; R: Report; D: Demonstrator; O: Other; PU: Public; PP: Shared only Among Participants

Table 1.3 c: List of milestones

Milestone Number	Milestone Name	Work Package(s) Involved	Expected Date	Means of Verification
MS1	Prototype Modelling Software for Testing	WP7/8	M6	delivery to SMEs
MS2	Database Infrastructure, Query & Visualization Tools	WP1/2	M12	delivery to partners
MS3	Whole-cell Models Required for Biotechnological Applications	WP4	M24	delivery to SMEs
MS4	Finished Software Package for the Systems-Biology Modelling Cycle	WP3/7/8	M36	delivery to the public
MS5	Proof-of-Principle Models Developed Using our Software	WP4–7	M36	publication/reports

Table 1.3d: Work Package Description

Work package number	WP1		**Start date or starting event:**		M1		
Work package title	Database Integration & Exploitation						
Activity type	RTD						
Participant number	1	3	4	**6**	7	8	9
Participant short name	CRG	EMBL	UvA	**FTELE.IGM**	UNIMAN	USheff	CSM
Person-months per participant:	3	12	3	**12**	6	3	2

Objectives

To develop new software tools and workflows for semi-automated integration and exploitation of diverse genomics, network and expression databases for model building.

Description of work, and role of participants

Task 1.1: Development of a database (NetBase) compliant with data standards (MIAME, MIAPE etc.) to store experimental data/meta-data and literature-derived knowledge in a standardised "computationally-ready" format to be easily used by visualisation and modelling tools developed in the course of the project. The database will cover multiple organisms, including human, mouse, *Drosophila*, yeast and *E. coli* (FTELE.IGM, EMBL).

Task 1.2: To connect the database infrastructure with the other tools used and developed in the consortium, in particular, DataRail (to process and visualise data), CellNOpt (for logical modelling; we will use networks generated from NetBase as prior knowledge), and the modelling/optimisation tools to be developed in WP3 of this project (FTELE.IGM, EMBL).

Task 1.3: To integrate disparate data sources (such as ChIP-seq and microarray gene expression data) through probabilistic models (USheff).

Task 1.4: To develop standards and tools for integration and comparison of spatial gene expression data within and between species (CRG, UvA).

Task 1.5: To integrate the tools developed in Tasks 1–3 into a common software framework, suitable for biotechnological applications (CSM).

Task 1.6: To integrate data from metabolomics experiments and flux balance analysis from *E. coli* (UNIMAN).

Deliverables (brief description and month of delivery)

D1.1: A relational database infrastructure named NetBase to be developed at FTELE.IGM for the purpose of integrating interaction and expression data from diverse data sources (M6).

D1.2: Interfaces with NetBase and CellNOpt to transfer prior knowledge networks (M6).

D1.3: Adaptable workflows for modellers to put together datasets for model building and fitting in a flexible and user-friendly way, implemented in a unified software framework (M12).

D1.4: Command-line and graphical software tools to create and manage database integration workflows (M12).

D1.5: An editor, which allows the user to create data files for modelling and enables auto-mated consistency and completeness checks (M18).

Work package number	WP2	**Start date or starting event:**		M1
Work package title	Visualisation Tools for Data Analysis & Model Building			
Activity type	RTD			
Participant number	1	3	4	**8**	9
Participant short name	CRG	EMBL	UvA	**USheff**	CSM
Person-months per participant:	3	3	3	**8.2**	3

Objectives

To develop new visualisation methods and tools to aid modellers in identifying relevant features, clusters and trends in the data, to identify relevant systems components, and to analyse highly complex non-linear network models.

Description of work, and role of participants

Task 2.1 We will use probabilistic, dynamical, latent variable models, for jointly visualizing disparate high-dimensional data sources. In particular these will be based on Gaussian process models (GPLVM) for data visualization originally introduced by Lawrence (UShef).

Task 2.2: Extension of DataRail visualisation routines for multi-dimensional data to be applied to the type of data used in the consortium, and integration with GPLVM and other tools (CSM, EMBL, in collaboration with the Sorger Lab at Harvard).

Task 2.3: We will develop tools (based on existing code, implemented in Java and Python) to systematically analyze and compare spatial gene expression patterns (CRG, UvA).

Deliverables (brief description and month of delivery)

D2.1: Software implementation of GPLVM with extended capability for visualisation of high-dimensional, heteroscedastic data with time-series structure (M12).

D2.2: Interface for DataRail and other tools developed in the consortium (in particular, GPLVM). Extended DataRail routines for visualisation (M12).

D2.3: Visualisation and comparison tools for spatial gene expression patterns (M12).

Work package number	WP3	**Start date or starting event:**		M1
Work package title	Integrated Software Tools for the Modelling Cycle			
Activity type	RTD			
Participant number	1	**2**	3	4	5
Participant short name	CRG	**CSIC**	EMBL	UvA	CWI
Person-months per participant:	18	**28**	3	15	30
Participant number	6	7	8	9	
Participant short name	FTELE.IGM	UNIMAN	USheff	CSM	
Person-months per participant:	12	6	17	10	

Objectives

To develop novel methods to support the model building cycle, and to integrate them into a unified, powerful and easy-to-use software framework, which can be applied to a wide range of modelling activities and processes.

Description of work, and role of participants

Task 3.1: We will implement Bayesian approaches to model building for tractable models based on differential equations and Gaussian processes, as well as for less tractable models based on non-linear differential equations and probabilistic modelling where Markov Chain Monte Carlo methods are required for parameter inference (USheff, FTELE.IGM)

Task 3.2: We will develop new parameter estimation strategies, based on stochastic global optimisation algorithms. These will be paired with fast local search algorithms to yield powerful hybrid search strategies (CSIC, CRG, UNIMAN).

Task 3.3: We will implement parallel meta-heuristics, which automatically favour specific optimisation strategies developed in T3.2 according to the measured current efficiency of each algorithm. These techniques will be implemented in software toolboxes which allow the user to choose among a wide range of powerful global search methods, taking advantage of parallel high-performance computers (including GPU-based architectures), as well as distributed/cloud computing on variable architectures (CSIC, CRG, CWI, UNIMAN).

Task 3.4: We will develop efficient algorithms for parameter estimation via multi-objective optimisation (for example, maximising both goodness of fit and robustness of the resulting network models). (CSIC, CRG, UvA, CWI).

Task 3.5: Development of novel methods, protocols and software tools for model building, with a special focus on multi-scale modelling, model selection and discrimination, parameter identifiability analysis (both theoretical and practical), model validation and uncertainty quantification (CWI, CSIC, CSM, INSIL).

Task 3.6: Integration of the above methods with the CellNOpt platform for large-scale logic modelling (EMBL).

Deliverables (brief description and month of delivery)

D3.1: New algorithms based on a Bayesian approach to mutual information to identify genome-wide regulatory network topologies from heterogeneous information (M18).

D3.2: New software tools for parameter estimation via global non-linear optimisation (including co-operative parallel meta-heuristics making use of high performance computing facilities (incl. GPU-based architectures) (M18).

D3.3: New software tools for multi-objective optimisation, implementing a wide range of cost functions (M24).

D3.4: Integrated software-suite for iterative multi-scale model building providing tools for all the steps in the modelling cycle; documentation describing the suite, incl. algorithm comparison & applications (M36).

Work package number	WP4	**Start date or starting event:**		M1	
Work package title	Application: Large-scale Models of Microorganisms and Eukaryotic Cell Lines				
Activity type	RTD				
Participant number	2	5	6	**7**	10	11
Participant short name	CSIC	CWI	FTELE.IGM	**UNIMAN**	INSIL	FS
Person-months per participant:	3	1	4	23	10	6

Objective

To apply the methods and software tools developed in WP1–3 for reconstructing and verifying large-scale models of metabolism and gene regulation.

Description of work

Task 4.1: Adopt and improve existing whole-genome metabolic reconstructions of *E. coli* and *S. cervisiae* in terms of annotation standards for further use in dynamic modelling (UNIMAN, FS, INSIL)

Task 4.2: Develop approximate kinetic models based on the *E. coli* and *S. cervisae* reconstructions using generic kinetic rate laws (lin-log, convenience kinetics or others). (CSIC, UNIMAN, INSIL, FS)

Task 4.3: Reverse-engineering of gene regulatory network of *E. coli* using publicly available transcriptomics data (microarrays, next-gen sequencing, etc.) (FTELE.IGM, CWI)

Task 4.4: Connection of gene regulatory network with metabolic network to create a multi-scale model of *E. coli* (UNIMAN, FTELE.IGM)

Task 4.5: Adopt the large-scale Chinese Hamster Ovary (CHO) cell metabolism reconstruction contributed by Insilico Biotechnology and update it according to established annotation standards. (UNIMAN, INSIL)

Task 4.6: Develop kinetic models for CHO cell metabolism based on the reconstruction and data existing at Insilico Biotechnology, and using generic kinetic rate laws and parameter estimation. (CSIC, UNIMAN, INSIL)

Deliverables (brief description and month of delivery)

D4.1: Standardized network reconstruction of *E. coli* metabolism expressed in SBML (M6).

D4.2: Genome-scale kinetic metabolic model of *S. cerevisiae* (for WP7) (M12).

D4.3: Standardized network reconstruction of CHO cell metabolism in SBML (M12).

D4.4: Genome-scale kinetic metabolic model of *E. coli* expressed in SBML (M18).

D4.5: Gene regulatory network of *E. coli* expressed in SBML (M24).

D4.6: Combined metabolic and genetic regulation model of *E. coli* (M36).

Work package number	WP5	**Start date or starting event:**		M12	
Work package title	Application: Signalling and Regulatory Networks in Cells				
Activity type	RTD				
Participant number	2	**3**	5	6	8	10
Participant short name	CSIC	**EMBL**	CWI	FTELE.IGM	USheff	INSIL
Person-months per participant:	3	**18**	2	8	9	6

Objective

To apply the methods and software tools developed in WP1–3 to models of signalling and regulatory networks in cell lines, and then link these models to the metabolic models of CHO cells developed in WP4.

Description of work, and role of participants

Task 5.1: Reconstruction of networks of signal transduction and gene regulation of relevance in biotechnological production processes, based on methods and data resources from WP1–3 (EMBL, FTELE.IGM, USheff).

Task 5.2: Calibration of network models using methods implementing the modelling cycle as described in WP3 (EMBL, CSIC, CWI, FTELE.IGM, USheff).

Task 5.3: Analysis of models to gain mechanistic and predictive insights into optimisation of biotechnological production processes (EMBL, FTELE.IGM).

Task 5.4: To link these models to models of metabolism in CHO cells developed in WP4 (EMBL, INSIL).

Deliverables (brief description and month of delivery)

D5.1: Algorithms that allow the integration of protein and gene expression measurements to obtain a hypothesized set of interactions for relevant signalling cascades (M18).

D5.2: A reconstruction of signalling and regulatory networks of relevance the production of nutraceuticals and other components in CHO cells in SBML format (M24).

D5.3: Kinetic models of signalling and regulatory networks in CHO cells (M30).

D5.4: Integrated models of signalling, regulatory, and metabolic networks in CHO cells (M36).

Work package number	WP6	**Start date or starting event:**	M12
Work package title	Application: Developmental Gene Regulatory Networks		
Activity type	RTD		
Participant number	1	**4**	5	8
Participant short name	CRG	**UvA**	CWI	USheff
Person-months per participant:	24	18.6	2	6

Objective

To apply the methods and software tools developed in WP1–3 to complex spatial models of gene regulatory networks involved in animal development.

Description of work, and role of participants

Task 6.1: Data integration/visualisation tools from WP1 & 2 will be used to quantitatively compare spatial gene expression data from different databases within species (e.g. mRNA vs protein data), and between species (e.g. different species of dipterans) (UvA, CRG).

Task 6.2: The datafile editor from WP1 will be used to create extended datasets for modelling spatial gene regulation in cnidarians (*Nematostella*) & *Drosophila* (UvA, CRG).

Task 6.3: The modelling cycle will be employed (using tools developed in WP3) to create new and improved models of developmental gene networks underlying pattern formation during the early development of *Nematostella* and various dipteran insects. Our software framework will allow a systematic comparison of optimisation algorithms and modelling frameworks for this problem, which is representative for many other complex spatial modelling applications in general (UvA, CRG, CWI, USheff).

Task 6.4: These models will be analysed to gain new biological insights into the pattern forming processes underlying animal form, and their evolution (UvA, CRG).

Deliverables (brief description and month of delivery)

D6.1: Improved/standardised datasets of spatial gene expression during animal develop-ment (M18).

D.6.2: Improved models of gene regulatory networks underlying pattern formation in animal development; datasets, models and biological analyses to be described in a number of separate manuscripts (M36).

Work package number	WP7	**Start date or starting event:**		M1
Work package title	Application: Validation of Network Models in Biotechnological Production Processes			
Activity type	RTD			
Participant number	4	5	7	10	**11**
Participant short name	CSIC	CWI	UNIMAN	INSIL	**FS**
Person-months per participant:	3	1	1	20	**30**

Objective

To apply the methods and software tools developed in WP1–3 to production processes in industrial biotechnology, with a strong focus on validating code and software for the application of fungal, bacterial and mammalian models (including dynamic models of *S.cerevisiae, E.coli* and CHO cell cultures). Areas of application are the optimization of industrial processes focusing on the production of nutraceutical ingredients, biopharmaceuticals, and fine chemicals.

Description of work, and role of participants

Task 7.1: We will provide recommendations for software design suitable for use in a commercial biotechnology setting (functionality and GUI) (CWI, UNIMAN, INSIL, FS).

Task 7.2: We will test tools/algorithms developed in WP1–3 and models developed in WP4. Scientists with different levels of modelling experiences (molecular biologist, engineer, bioinformatician) will be employed as testers. Feedback for the improvement of the software will be provided to CSM and the academic partners (INSIL, FS)

Task 7.3: We will develop simulations (through iterative application of the modelling cycle; based on models from WP4) of the production of nutraceutical ingredients, pharmaceuticals, and fine chemicals. This will include integration of published data on transcription and metabolism, as well as results obtained in WP4/5 (CSIC, INSIL, FS).

Task 7.4: We will use the models developed in Task 7.3 to compare low-, medium- & high-producer strains. Model results will be used to describe phenotypic data and to identify metabolic engineering targets, as well as targets for process improvement (INSIL, FS).

Deliverables (brief description and month of delivery)

D7.1: Recommendations for software design (functionality and GUI) (M3–24).

D7.2: User-friendly version of prototype software for testing in a setting for industrial applications (M6–24).

D7.3: Models (based on WP4) for simulation of the production of nutraceutical ingredients, pharmaceuticals, and fine chemicals in microorganisms and eukaryotic cell lines (M24).

D7.4: Comparative analysis of low-, medium-, high-producer strains using software developed by partners (integration of data at flux, transcript and metabolite level) (M30).

D7.5: Targets for metabolic engineering, synthetic biology and process improvement for various production organisms including *E.coli, S.cerevisiae* and CHO cells (M36).

Work package number	WP8	**Start date or starting event:**		M1			
Work package title	Dissemination & Exploitation						
Activity type	OTHER & DEM						
Participant number	1	**9**					
Participant short name	CRG	**CSM**					
Person-months per participant:	4	**19**					

Objectives

To disseminate project achievements, to implement a sustainable, permanent distribution and support base of our software, to advertise this software at scientific meetings and professional trade shows, and to explore commercialization of project results. All of the above will be done in accordance with the regulations of the CA and will in particular be subject to the prior conclusion of written agreements with the respective institutes.

Description of work, and role of participants

Task 8.1: We will disseminate main project achievements through the central project website, peer-review publications and press releases to the media (CRG).

Task 8.2: We will set up a version-control server for source code (SVN), as well as automatic building and testing processes (CMake.org) with web-based reporting (CDash.org). These efforts will be co-ordinated by partner 9 (CSM), who will offer their existing code development infrastructure to the project, and will provide training in its use for the other partners (see WP9).

Task 8.3: We will create a unified and consistent cross-platform code infrastructure (integrating efficient, native numerical code in C/C++ with graphical user interfaces based on the Tulip widget library) that includes all the methods and tools implemented and developed during this project, which enables the easy establishment of flexible, automated workflows, and guarantees interoperability and comparison of methods and tools (CSM).

Task 8.4: We will present our software at selected scientific meetings, and relevant professional trade shows (for example, the Annual International Conference on Intelligent Systems for Molecular Biology (ISMB); the International Conference on Systems Biology (ICSB); the RECOMB Conference with its DREAM Initiative for optimisation algorithms; the European Conference on Computational Biology (ECCB); the symposium on Computer Applications in Biotechnology to be held in 2013; and the annual Bio-IT World Conference & Expo). This will be done by means of oral presentations and posters, as well as demonstration stalls, where potential users can interactively explore our software, and where they will be provided with professional advice, instructional material and documentation (CSM, CRG).

Deliverables (brief description and month of delivery)

D8.1: Project website for the scientific community and the general public (M6).

D8.2: Software development and testing architecture (version-control server for code, automated building/testing processes with web-based reporting; hosted by CSM) (M6).

D8.3: Integrated software suite implementing the methods and tools developed during this project in an interoperable, user-friendly and well-supported way (M36).

D8.4: Oral/presentations/demo stalls at selected scientific conferences and relevant professional trade shows (M18–36).

D8.5: Peer-reviewed publications in high-profile journal describing our software tools, accompanied by press releases to the media wherever possible/appropriate (M18–36).

Work package number	WP9	**Start date or starting event:**		M1			
Work package title	Training						
Activity type	OTHER						
Participant number	**1**	9					
Participant short name	**CRG**	CSM					
Person-months per participant:	**4**	2					

Objectives

To provide multi-disciplinary training to post-doctoral research fellows and other researchers (such as PhD students) in our groups and companies in the state-of-the-art methods of our fields, and to train researchers not directly involved in our consortium in the methods and tools which are to be developed during this project, and consequently in the effective interpretation and use of scientific data.

Description of work, and role of participants

Task 9.1: Two week-long workshops will be organised, at the CRG and at the EBI/EMBL, early and late in the project. The early workshop will aim at teaching post-doctoral fellows and other researchers involved in our consortium how to understand and apply state-of-the-art methods and tools involved in database integration, visualisation and model building (taught/organized jointly by all academic partners; CSM will provide training for their code development and modelling frameworks). In addition, this first workshop will include a session on intellectual property, technology transfer, and entrepreneurship (organized by the Innovation Board; see section 2.1). The second workshop, towards the end of the project, will aim at teaching outside researchers the theoretical and practical aspects of the methods and tools developed during this project (teaching co-ordinated by CSM; involving all academic partners in the network).

Task 9.2: In addition, we plan to encourage PhD students and post-doctoral researchers involved in our consortium to swap laboratories during the execution of their projects. The idea is to have researchers spend one or two years in one lab, before moving on to another partner of the consortium. In this way, we ensure that young researchers are exposed to many different, but related fields of expertise, such that they can combine the skills they learn in new and productive ways in the future.

Task 9.3: We also want to foster exchanges of expertise and know-how between the academic and private groups through short or medium-term secondments of the fellows involved in the project.

Deliverables (brief description and month of delivery)

D9.1: Project-internal workshop at the CRG to train researchers within our consortium in state-of-the-art methods and tools in our respective fields of expertise (M12).

D9.2: Publicly advertised workshop at the EBI/EMBL to train researchers in the general field of optimisation and modelling how to understand and apply the methods and software tools developed during this project (M36).

D9.3: Exchange visits and secondments between the academic and private participants (M1–36).

Work package number	WP10	**Start date or starting event:**		M1			
Work package title	Project Management						
Activity type	MGT						
Participant number	**1**						
Participant short name	**CRG**						
Person-months per participant:	**10**						

Objectives

To efficiently manage the project to ensure successful completion of the scientific and technological objectives within the planned time frame and budget and at high quality standards.

Description of work (all tasks will be carried out by the CRG)

Task 10.1: Consortium Management: This task covers the day-to-day management of the project, including the organisation of project meetings and events, the provision of a decision-making structure, conflict resolution and risk management.

Task 10.2: Quality Assurance: A Quality Assurance Plan will be defined to ensure consistency across the WPs and to guarantee that all deliverables have a high quality.

Task 10.3: Communication: We will ensure effective communication within the consortium and between the project and the EC by providing and implementing pertinent tools and mechanisms (website, mailing lists, phone conferences, reports, newsletters, etc).

Task 10.4: Financial and Legal Management: This task covers the management of EC payments to the partners, overview of budget expenditure, grant amendments, support to the partners in all financial and legal aspects to make sure that the requirements of the grant agreement are understood and fulfilled by the consortium members.

Task 10.5: Reporting: This task consists in gathering reports and deliverables from the WP leaders, and submitting them on behalf of the consortium to the EC.

Task 10.6: Monitoring ethics and gender issues: This task includes the monitoring and reviewing of any ethical issues identified in the proposal, as well as defining, implementing and monitoring actions to promote the participation of women in the project.

Deliverables (brief description and month of delivery)

D10.1: Signature of Consortium Agreement (M3)

D10.2: Establishment of Quality Assurance Plan (M3)

D10.3: Kick-off meeting (M3)

D10.4: 1st Annual Meeting (M14)

D10.5: 1st Annual Periodic Activity and Management Report (M14)

D10.6: 2nd Annual Meeting (M26)

D10.7: 2nd Annual Periodic Activity and Management Report (M26)

D10.8: Final Meeting (M36)

D10.9: 3rd Final Periodic Activity and Management Reports (M36)

Table 1.3 e: Summary of Staff Effort

Participant No./Short Name	WP 1	WP 2	WP 3	WP 4	WP 5	WP 6	WP 7	WP 8	WP 9	WP 10	Total Person Months
1. CRG	3	3	18			24		4*	4*	10*	66
2. CSIC			28	3	3		3				37
3. EMBL	12	3	3		18						36
4. UvA	3	3	15			18.6					39.6
5. CWI			30	1	2	2	1				36
6. FTELE.IGM	12		12	4	8						36
7. UNIMAN	6		6	23			1				36
8. USheff	3	8.2	17		9	6					43.2
9. CSM	2	3	10					19	2		36
10.INSIL				10	6		20				36
11.FS				6			30				36
Total	41	20.2	139	47	46	50.6	55	23	6	10	437.8

(*) 18 person-months in WP8–10 for a part-time project manager at the CRG (see section 2.4).

iv) Provide a Graphical Presentation of the Components Showing Their Inter-dependencies (Pert Diagram or Similar)

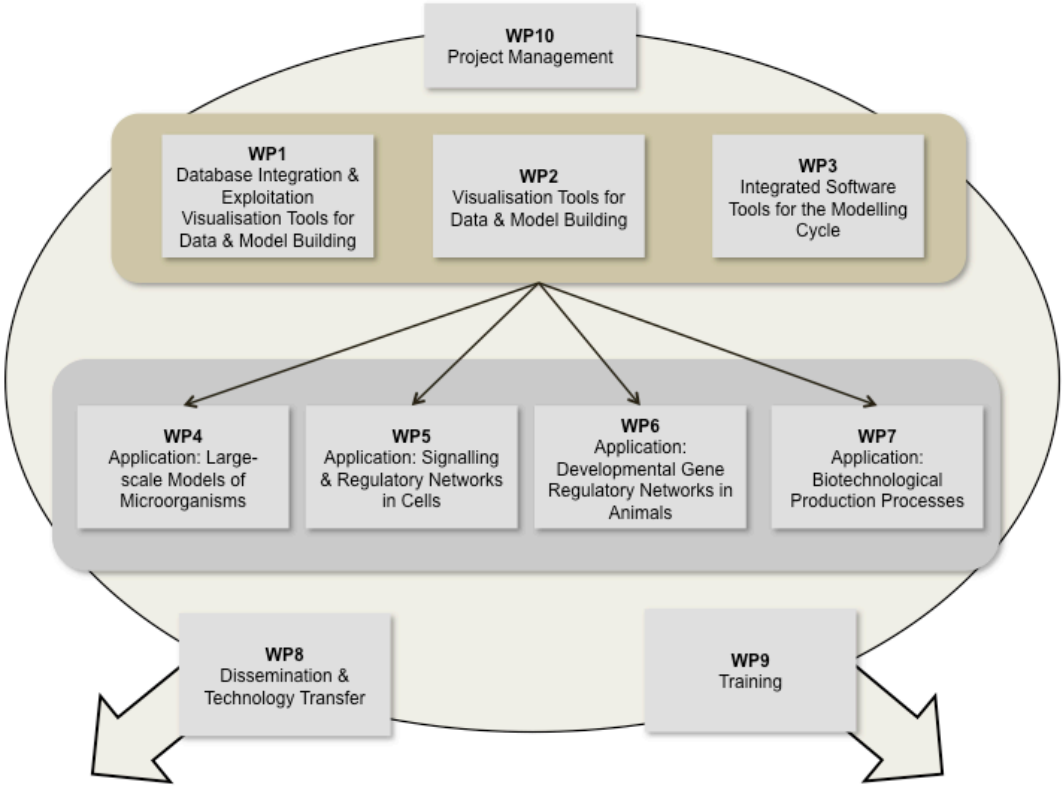

Figure 6: Work Packages of BioPreDyn and their Interdependencies. WP1–3 focus on method development (vertical objectives) while WP4–7 deal with the application of these methods to different biological and biotechnological benchmark problems. WP8 and WP9 are concerning with dissemination of results/products and training. Management of the entire project is the concern of WP10.

v) Describe any Significant Risks, and Associated Contingency Plans.

Work Packages 1 & 2 (WP1 & WP2):

We do not foresee any major risks for these work packages—mainly concerned with software design and data integration—due to the well-established methodologies we will use to develop the relational database as well as the query and visualisation tools. The main challenge relating to WP2 (visualisation) is the large-scale and diverse nature of the data. For risks associated with integrating large amounts of code from diverse sources (D1.4, D2.3) see WP8.

Work Package 3 (WP3):

There is always a risk of failure associated with developing novel numerical methods for non-linear modelling and optimisation. However, the synergistic and complementary expertise that we accumulate within our consortium will ensure that algorithm development will be up to the most stringent quality standards possible. D3.1 will start by considering the most tractable models, and methods will be extended and refined stepwise to deal with more challenging models based on non-linear differential equations and probabilistic frameworks, as well as with multi-objective optimisation (D3.2 and D3.3). We will pay particular attention to scalability issues, and high-performance computing will be used to keep them under control. Issues with parameter identifiability will be addressed from a practical point of view with suitable nonparametric statistical tools. For risks associated with large-scale coding projects (D3.4) see WP8.

Work Package 4 (WP4):

D4.1 relies on published work, and is therefore of low risk. D4.2 and D4.4 involve the inference of a large number of parameters based on existing data. This task will be informed by the methodologies developed in WP3 and we do not expect there to be any major issues. If our reverse-engineering methods fail, we can apply alternative reverse-engineering methods available from the literature. D4.3 is quite similar to D4.1. It is feasible since partner INSIL already have a CHO cell metabolism map. However, it involves a slightly higher risk of failure than D4.1 because mammalian cells have a more complex metabolism than *S. cerevisiae*. D4.6 is the most risky activity of this work package. We may be unable to construct the large kinetic model required for such an integrated reconstruction of a cell. This may be due to lack of data: if we are unable to obtain enough data to constrain the model, we will proceed with an reduced model based on flux distribution (i.e. containing a smaller number of pathways, those that carry most of the flux); such a model would still be of value in biotechnological applications.

Work Package 5 (WP5):

D5.1 is low-risk: it relies on previous work of partners and other groups on data integration. The technological requirements for D5.2 and D5.3 are provided by WP1–3. Potential bottlenecks include the availability of data on signalling and regulatory processes in CHO cells. If available data re not sufficient to construct models, we will utilise data from related cell types, to create a model resembling CHO cells as closely as possible. Finally, D5.4. is a very challenging deliverable; if a fully integrated model of metabolic, signalling, and regulatory networks is not achieved, we are confident to provide at least specific models for the different regulatory scales and processes.

Work Package 6 (WP6):

It is notoriously difficult to standardise spatial gene expression data (D6.1), due to difficulties in comparing developmental stages and types of tissues across species. However, this is not a serious problem for our experimental systems, since dipteran (fly) embryos are morphologically very similar, while outside the dipteran system we can fall back on qualitative comparisons should more rigorous standardisation efforts fail. For D6.2, we do not foresee any major risks for modelling pattern formation in flies, since a proof-of-principle that reverse engineering in this system works is already available (unpublished). In *Nematostella*, the major risk is that the available data may not be sufficient to constrain the fitting problem, and we may not be able to obtain unique solutions for our fits. This will provide a challenge for algorithms concerned with parameter identifiability and optimal experimental design from WP3, which are designed to address such problems.

Work Package 7 (WP7):
Metabolic target identification (D7.5) can be challenging and thorough model validation needs to be conducted using known established targets. Furthermore, the aim of WP7 is to establish novel targets for improvement of biotechnology production processes. Here it is essential that sufficient experimental data is available. Should this not be the case, FS is willing to generate data outside BioPreDyn in order to have sufficient data available for simulation.

Work Packages 8 (WP8):

While dissemination and exploitation of results does not require risk assessment, this package also contains deliverables based on large-scale coding projects (D8.2, D8.3). The tasks underlying these deliverables are designed to handle the complexities of such efforts. Collaborative coding practices and version control will be ensured by a SVN server to which all partners will have access. Furthermore, automatic code-building and -testing tools (cmake.org) with web-based reporting (cdash.org) will be set up. The code integration part of this project will be co-ordinated by CSM whose personnel has ample experience with such large-scale collaborative coding projects.

Work Packages 9–10 (WP9–10):

Risk assessment is not necessary for these work packages since they do not deal with research- and technology-related aspects. For management procedures (WP10) see section 2.1.

2. Implementation

2.1 Management Structure and Procedures

BioPreDyn is a multidisciplinary project that brings together eight academic institutions and three SMEs in seven different European countries. Due to the complexity and interdisciplinary/sectorial nature of the project (and due to the fact that in basic research events can sometimes take unexpected turns), we will establish effective management structures and procedures from the very beginning. Our management strategy will allow continuous monitoring of the project, taking timely corrective actions whenever needed, sharing resources and technologies for a synergistic outcome, and protecting, publishing, and utilising the knowledge generated. All the partners will agree on management structures and procedures, which will be illustrated in detail in the **Consortium Agreement**.

2.1.1 Management structure

Fig. 7 illustrates the management structure of BioPreDyn, including the main players and their relationships.

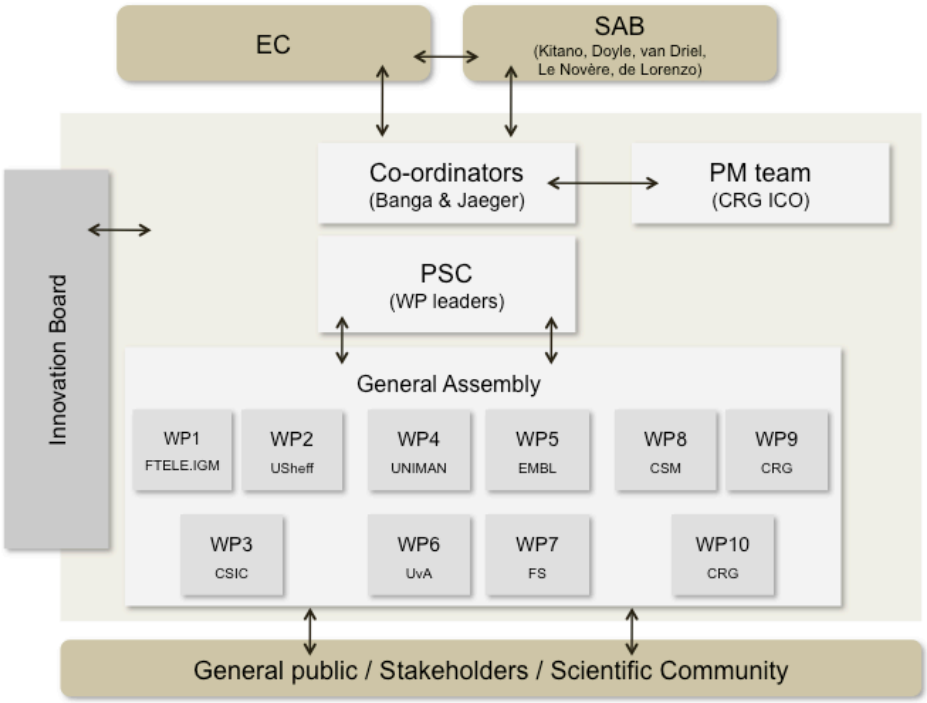

Figure 7: Management Structure of BioPreDyn

Scientific Co-ordinators

Dr. Banga and Dr. Jaeger will take charge, in a synergistic fashion, of the scientific co-ordination of the project. Dr. Banga will mainly supervise the activities related to tool/algorithm development and Dr. Jaeger will mainly supervise the activities related to their applications in research and biotechnology. Their role includes acting as intermediary between the consortium and the European Commission (EC), as well as chairing the Project Steering Committee and the General Assembly. They will interact weekly with the Project Manager (PM) to ensure the success of the project within its defined budget and time-period.

Neither Dr. Banga nor Dr. Jaeger has previous experience with co-ordinating European Framework programs. However, both of them have been involved in European projects as partners (Dr. Banga: 7 projects, 3 of them in the field of systems biology; Dr. Jaeger: 2 projects, both sponsored by the EraNet initiative, concerned with optimisation and modelling), and they will be supported by the CRG International Collaboration Office (ICO; see next section) in their role as project co-ordinators.

Project Management (PM) Team

Management activities will be performed by the CRG. A dedicated project manager with suitable administrative skills as well as scientific background will be hired to manage BioPreDyn successfully, and provide day-by-day assistance to the scientific co-ordinators and the partners. The project manager will be incorporated into the CRG International Collaboration Office (ICO; headed by Dr. Michela Bertero), which has long-standing and extensive experience in successful management of European collaborative projects (both FP6 and FP7). The ICO works in tight collaboration with the CRG Research, Legal, Communication and Technology Transfer Offices. Apart from assisting the Scientific Co-ordinators, the project manager will be responsible for the following tasks:

- preparation of the Consortium Agreement,
- co-ordination of all contractual issues,
- preparation and timely submission of deliverables, reports and financial statements,
- monitoring of budget use, and distribution of funds to the partners,
- streamline communication flows within the consortium, as well as with the external scientific community and the general public,
- provide support for the organisation of project meetings, workshops, phone conferences, and other events,
- supervise gender and ethical issues,
- oversee and support the activities of the different project committees.

Project Steering Committee (PSC)

The PSC will be formed by the eight leaders of work packages 1–8, and will be chaired by the two scientific co-ordinators (Drs. Jaeger and Banga). Meetings (via conference call or face-to-face) will be held on a regular basis every 3 months. The PSC will have the following tasks:

- strategic decisions concerning the scientific and technological activities and allocation/distribution of funds,
- ensuring that there is an effective communication flow between partners and between the consortium and the EC,
- resolving conflicts among partners and project committees,
- preparing topics of discussions for the General Assembly (GA),
- implementing technical and scientific details of the work plan, taking into account recommendations of the EC, the Scientific Advisory Board (SAB) and other project committees.

Work Package Leaders

Each work package (WP) will be supervised by one leader, as agreed during the preparation of this proposal. His/her responsibilities will include:

- supervision of the scientific and technological activities within the assigned work package, including identification of potential bottlenecks,
- reporting to the PSC and the co-ordinators.

General Assembly (GA)

The GA is the ultimate decision-making body of the consortium. It will be composed of one representative from each partner to ensure that all views are represented in the decision-making process. The GA will meet at least once a year during the Annual Meetings, but extraordinary meetings may be convened by the PSC or the co-ordinators to address specific issues.

The GA will be chaired by the two co-ordinators and decide on all fundamental decisions for the project implementation such as:

- implementing changes in the overall project work plan, introducing new partners and re-allocation of tasks and budget,
- resolving conflicts, which could not be settled by the PSC,
- taking actions to be taken with regard to a defaulting party,
- deciding on changes to the Consortium Agreement.

Innovation Board (IB)

The IB will be appointed at the kick-off meeting and will be central to the dedicated work package activities. Members of the IB will include representatives from all three SMEs, experts in technology transfer from partner institutes, and experts in software licensing. The IB's main tasks will include:

- evaluate the licenses linked to background software and databases to ensure that foreground software and databases to be developed are free of unwanted restrictions for the final aims of use, distribution and exploitation,
- indentify discoveries and inventions with commercial potential,
- provide consultancy to the partners on the feasibility and the procedure for protecting and exploiting the knowledge generated by the project,
- help seek (where necessary) industrial partners for further commercialization,
- assist in the stipulation of confidentiality and understanding agreements with external partners, and
- mentor BioPreDyn researchers to broaden their career perspectives in the private sector.

Scientific Advisory Board (SAB)

A Scientific Advisory Board will be appointed at the kick-off meeting and will have the aim to assess the progress and quality of the work carried out by the consortium, and further to provide advice on the scientific directions of BioPreDyn. The SAB will be invited to the Annual Meetings and will receive Annual and Interim Reports in advance. It will be composed of renowned scientists from academic institutions and industry.

Very high-profile scientists (such as Hiroaki Kitano, Francis Doyle, Roel van Driel, Nicolas le Novère, and Victor de Lorenzo) have already agreed to serve on the SAB if the project is positively evaluated.

2.1.2 Management procedures

The following management procedures will provide the adequate framework for an efficient and smooth implementation of the project.

Consortium Agreement

The consortium members will negotiate, agree and sign a **Consortium Agreement** before the start of the project based on the DESCA model contract. The Consortium Agreement will regulate issues related to management structure and procedures, quality control, communication, financial and legal aspects, decision-making and conflict resolution mechanisms, risk management, management of intellectual property, etc as summarized in the following paragraphs.

Quality Assurance

A crucial element of the management procedure of BioPreDyn will be a straightforward quality control system. The co-ordinators will be responsible for the production of the **Quality Assurance Plan**, which will include guidelines and references for good practices and whose implementation will be the joint responsibility of all partners. Quality needs to be controlled mostly at three levels: 1) generation of new software tools; 2) increasing interpretative and predictive capacity of data generated; and 3) testing and application of computational models generated during the project. The WP leaders will be responsible for quality control and, together with the PSC, to identify promptly any risk, delay or other factors that might affect the work plan.

Communication Management

The PM team will set up effective tools for the efficient and transparent flow of communication among project partners. These tools include mailing lists, website and intranet, phone conferences, interim reports and newsletters. The following **mailing lists** will be set up: a general mailing list for all consortium members, specific mailing lists for the steering committee, administrative issues, and others whenever needed. In addition to the public part for visibility of the project (see section 3.2 for further details), the **webpage** will host a secured **intranet** dedicated to deposit reports and contractual documents, to host a forum, and to exchange scientific material. **Phone conferences** will be organized every 3 months with PSC members or upon request. In addition to the official

reports to the EC, short **interim reports** will be prepared and shared with all partners every 6 months. These reports will allow monitoring as well as sharing results among the consortium members. A **BioPreDyn Newsletter** will be edited and distributed regularly to highlight major project achievements, news and upcoming project meetings or events.

Furthermore, the communication flow will be facilitated by attendance to **project meetings** (annual project meetings and/or smaller meetings involving subprojects or specific WPs) and **other related events** (scientific conferences, training activities, workshops, etc).

Financial and Legal Management/Official Reporting

The PM team, with support of the CRG Financial Office, will be responsible for receiving the payments from the EC and distributing the funds to the partners according to the agreed budget shares. The **financial management** also involves: monitoring budget expenditures by the partners to ensure the appropriate use of resources, suggesting correction measures whenever applicable, and providing support to consortium members in all aspects related to financial issues, including financial audits and reports. Additionally, the PM team will make sure that all **legal requirements** derived from the grant agreement and consortium agreement are understood by the partners and fulfilled by the consortium.

The co-ordinators and the PM team will be the link between the Consortium and the EC Project Officer in charge of the project and will ensure **official reporting** (including deliverables) to the EC, according to the timing established in the Grant Agreement and Annex I.

Decision-Making Structure, Conflict Resolution and Risk Management

The **decision-making structure** for BioPreDyn has two levels: the General Assembly (GA) and the Project Steering Committee (PSC). Decisions will be made by the GA or PSC according to the responsibilities set out in the Consortium Agreement and briefly described in Section 2.1.1. All partners will appoint one representative and one deputy to the GA. Each member of the GA will have one vote. Decisions will ideally be made on the basis of consensus. If consensus cannot be achieved, they will be made on the basis of a majority vote with the co-ordinators having a casting vote. A quorum of 2/3 of the partners should be present or represented at the meeting.

The partners commit themselves to resolve any **conflict** amicably and as speedily as possible. Potential conflicts should be identified, analysed and resolved at the lowest level first (WP level). If the conflict cannot be solved at these levels between the partners concerned, the PSC will have both the responsibility and authority for conflict resolution as will be clearly defined in the Consortium Agreement.

Technical and scientific risks have been indentified in each work package (see section 1.3v). A procedure for **risk management** will be set out in the Consortium Agreement. Following this procedure, partners will be responsible for reporting (to the WP leaders, the PSC or the co-ordinators) any risks that might occur during the project lifetime and that might affect the successful completion of the project objectives. Depending on the risks identified and their impact on the project, the PSC or GA might be responsible to take corrective actions.

Management of Intellectual Property (IP)

The project will likely produce IP that is of significant value for the scientific community as well as for industrial partners (not restricted to the ones in the project). The project will maintain high awareness of opportunities to protect and exploit IP of potential commercial value, through a dedicated work package (WP8) and the establishment of the Innovation Board (IB). At the very beginning of the project, the IB will be in charge of assessing the status of existing licenses for each background software to be used in the project, and of evaluating and proposing concrete licensing strategies for the shared code to be developed during this project. Management of IP will be extensively described in the Consortium Agreement, which will be signed at the beginning of the project by all partners (see also section 3.2).

2.2 Individual Participants

Partner 1: CRG

Description:

The Centre for Genomic Regulation (CRG) is an emerging first-class research centre created in 2000 by the Catalan government and the University Pompeu Fabra (UPF) in Barcelona. The CRG's aim is to promote research excellence in biology and biomedicine. It provides an interdisciplinary and dynamic environment, in which researchers tackle a wide range of fundamental problems using 'omics' and systems-level approaches. The applicant's laboratory is part of the EMBL/CRG Research Unit in Systems Biology (Co-ordinator: Dr. Luis Serrano), a joint programme between the CRG and the European Molecular Biology Laboratory (EMBL). The CRG has extensive experience in co-ordinating European research projects, demonstrated by the fact that it is currently in charge of managing 4 such grants, and has previously co-ordinated 3 more projects under FP6.

Role in the Project:

Data integration and visualisation; parameter estimation, global optimisation algorithms; application: developmental gene regulatory networks in dipteran insects.

Expertise:

Our group is applying a reverse-engineering approach to the study of network evolution. We focus on the investigation of pattern-forming networks active during development of dipteran insects (flies, midges and mosquitoes). Our main model system—the gap gene network involved in segment determination during early development—will serve as one of the test cases for the reverse-engineering methods during this project. It is an ideal network to study in this context, since it represents a typical developmental gene regulatory network with a moderate number of components, but high spatial and temporal regulatory complexity. Comprehensive, quantitative datasets of spatial gap gene expression patterns are available. Our group has extensive expertise in data acquisition/quantification, global non-linear optimisation, and data/model analysis by means of graphical and numerical methods.

Selected recent publications (3 max):

Jostins L & Jaeger J (2010). Reverse engineering a gene network using an asynchronous parallel evolution strategy. *BMC Syst Biol* 4:17.

Ashyraliyev M, Siggens K, Janssens H, Blom J, Akam M & Jaeger J. Gene Circuit Analysis of the Terminal Gap Gene *huckebein*. *PLoS Comp Biol* 5: e10000548.

Jaeger J, Surkova S, Blagov M, Janssens H, Kosman D, Kozlov KN, Manu, Myasnikova E, Vanario-Alonso CE, Samsonova M, Sharp DH & Reinitz J (2004). Dynamic control of positional information in the early *Drosophila* blastoderm. *Nature* 430: 368–71.

Key Personnel:

Dr. Johannes Jaeger (PI) is a developmental geneticist, who has been trained in modelling and reverse-engineering during his MSc (with Prof. Brian Goodwin, 2000) and PhD (with Prof. John Reinitz, 2006). During his post-doc at the University Museum of Zoology in Cambridge (UK, supervisor: Prof. Michael Akam), and his time as a group leader at the CRG (from Oct, 2008), he has been applying quantitative, data-driven modelling approaches to the study of the developmental and evolutionary dynamics of gene regulatory networks.

Dr. Anton Crombach (post-doc) is a computer scientist by training, who did a PhD in the field of *in silico* evolution (with Prof. Paulien Hogeweg, Utrecht, NL). He is currently carrying out modelling/parameter estimation for gene network models, and evolutionary simulations.

Damjan Cicin-Sain is our group's programmer. He implements image processing and database tools, as well as high-performance code for model optimisation.

A post-doc, to be hired on this project, will be carrying out systematic comparisons of optimisation algorithms and modelling frameworks applied to the problem of pattern formation in early fly embryos.

Partner 2: CSIC

Description:

The Agencia Estatal Consejo Superior de Investigaciones Científicas (CSIC) is an autonomous, multi-disciplinary public research body affiliated to the Spanish Government. CSIC is the largest public research body in Spain, with its own legal structure and is represented throughout the Spanish territory with a total of 126 centres/institutes. The team participating in this project, the Bio-Process Engineering Group, is located at the Instituto de Investigaciones Marinas (IIM-CSIC) in Vigo, in the North-West of Spain. CSIC has considerable experience in both participating and managing R&D projects and training grants. Under the 7th Framework Programme, the CSIC has signed 129 actions (18 co-ordinated by the CSIC). The CSIC has been the 5th organisation in Europe in project execution and funding in the 6th Framework Programme.

Role in the Project:

Parameter estimation, global optimisation algorithms; model reduction, model selection and discrimination; parameter identifiability analysis; optimal experimental design.

Expertise:

The Bio-Process Engineering Group has strong expertise in dynamic modelling of biological systems, with emphasis on (i) robust parameter estimation of nonlinear dynamic models, and optimal experimental design, (ii) optimisation (local, global; single and multi-objective) and optimal control of bio-systems, (iii) model-based control, including robust and non-linear model predictive control (iv) sensitivity and identifiability analysis.

Selected recent publications (3 max):

Ross J, Villaverde AF, Banga JR, Vazquez S, Moran F (2010) A generalized Fisher equation and its utility in chemical kinetics. *Proc Natl Acad Sci USA* 107: 12777–81;

Balsa-Canto E, Alonso AA & Banga JR (2010). An iterative identification procedure for dynamic modeling of biochemical networks. *BMC Syst Biol* 4:11.

Banga JR & Balsa-Canto E (2008). Parameter estimation and optimal experimental design. *Essays in Biochemistry* 45:195–210.

Key Personnel:

Presently, the research activities of the Bio-Process Engineering Group are carried out by 15 persons: 3 permanent (tenured) scientists (Prof. Julio R. Banga, Dr. Antonio A. Alonso and Dr. Eva Balsa-Canto), plus a group of 7 PhD students and 5 post-docs.

Julio R. Banga is currently Research Professor of CSIC and leader of the BioProcess Engineering Group. He obtained a Ph.D. in Chemical Engineering from the University of Santiago de Compostela in 1991. During 1992, he was a post-doc at the University of California, Davis (USA), and after that he spent three years as Assist. Prof. of Chemical Engineering at the University of Vigo, Spain. During those years, he also spent periods as visiting researcher at the University of Pennsylvania and at the M.I.T. (USA). Since 1996, he is a tenured researcher at CSIC.

His main research topic is the application of mathematical modelling and optimisation to biological processes and systems, with applications targeting the areas of bioprocess engineering and systems biology. He has supervised over ten PhD students. He is the author of more than 110 archival publications, and has been involved in over 40 major research projects and contracts, including 4 EU projects in the area of systems biology. Currently, he is a member of the Editorial Board of BMC Systems Biology, a member of the IFAC Technical Committee on Control of Biotechnological Processes, and a member of several European external advisory boards. Dr Antonio A. Alonso is specialised in the analysis and control of nonlinear dynamic systems, with many applications in the bio-systems area (over 70 research papers). Dr. Balsa-Canto is an expert in systems identification and identifiability analysis.

Partner: EMBL

Description:

The European Molecular Biology Laboratory (EMBL) is a molecular biology research institution supported by 20 European countries and Australia as associate member state. The headquarters of EMBL are in Heidelberg, Germany. The team participating in this project, the Systems Biomedicine Group, is based at the European Bioinformatics Institute (EBI), an EMBL-outstation located at the Wellcome Trust Genome Campus, Hinxton, near Cambridge. EBI can be considered the European centre for globally co-ordinated efforts to collect and disseminate biological data (e.g. EMBL Nucleotide Sequence Database, UniProt, ArrayExpress, Ensembl, InterPro and BioModels) and over 180 other resources. As of March 2010 at the campus data centre, there are more than 8,000 cores of high performance computing in total and more than 7 Petabytes of raw disk. EBI provides state-of-the-art services to allow researchers to understand not only the molecular components that go towards constructing an organism, but how these parts combine to create systems. In addition, EMBL-EBI provides extensive scientific training for users of its services (e.g. 280 unique training-related events during 2008–2009).

Role in the Project:

Tools and methods to link models to experimental data; integration with data and network databases; modelling based on logical formalisms; applications to signalling networks.

Expertise:

The Systems Biomedicine Group has strong expertise in modelling signal transduction networks using logic formalisms and high-throughput proteomics data. The group develop methods and tools to leverage prior knowledge on biological networks from public sources with dedicated experimental data.

Selected recent publications (3 max):

Alexopoulos LG*, Saez-Rodriguez J*, Cosgrove B, Lauffenburger DA & Sorger PK (2010). Networks inferred from biochemical data reveal profound differences in TLR and inflammatory signaling between normal and transformed hepatocytes.
Mol Cell Proteomics 9: 1849.

Saez-Rodriguez J*, Alexopoulos LG*, Epperlein J, Samaga R, Lauffenburger DA, Klamt S & Sorger PK (2009). Discrete logic modeling as a means to link protein signaling networks with functional analysis of mammalian signal transduction. *Mol Syst Biol* 5: 331.

Saez-Rodriguez J*, Goldsipe A*, Muhlich J, Alexopoulos LG, Millard B, Lauffenburger DA, Sorger PK (2008). Flexible informatics for linking experimental data to mathematical models via DataRail. *Bioinformatics* 24:840–7.

[* denotes equal contribution]

Key Personnel:

Presently, the group consists of Julio Saez-Rodriguez (Principal Investigator), Jerry Wu (scientific programmer), and two post-doctoral fellows and one PhD student.

Julio Saez-Rodriguez studied Chemical Engineering in the Universities of Oviedo and Stuttgart (1996-2001, with distinctions from the Spanish Government), and performed his graduate studies at the Max-Planck-Institute for Dynamics of Complex Technical Systems (2002-2007); his PhD was awarded the MTZ-Award for the best Dissertation in Medical Systems Biology. From 2007 to 2010 he was a post-doctoral fellow at Harvard Medical School and M.I.T., in a project funded by Pfizer. He is since July 2010 a group leader at EMBL-EBI, with a joint appointment at the Genome Biology Unit in EMBL-Heidelberg, and a senior fellow in Wolfson College. He is also the co-organizer of the DREAM (Dialogues in Reverse Engineering assessment of methods) initiative. He has co-authored 22 papers in peer-reviewed international journals.

Partner 4: UvA

Description:

The Computational Science Group at the Universiteit van Amsterdam (UvA) seeks to discover, through modelling and simulation, the way distributed information is being processed in complex systems. We focus on theory, applications, and problem-solving environments. We address issues of how physical and biological problems can be formulated in this framework and how they can be mapped onto distributed computer architectures and grid systems. The applicability of this approach is validated through the development of high-performance distributed problem-solving environments for asynchronous natural processes. The group is proactive with respect to e-Science virtual laboratories. Its work has strong theoretical foundations together with tight couplings to biological applications. UvA has extensive experience in (the management of) EU Framework projects, including HPCNET, CrossGrid, ACGT, Morphex, COAST, ViroLab, QosCosGrid, MeDDiCa, and MAPPER among others.

Role in the Project:

Modelling gene regulatory networks, cell-based modelling, modelling and simulation of morphogenesis, optimisation algorithms, multi-objective optimisation.

Expertise:

Our group has long-standing experience in high-performance computing, scientific visualisation, modelling and simulation in computational biology, bio-medical applications and physics. Within computational biology we do research at a range of different levels of organisation (genome-gene regulatory networks-cells-tissue-organism). We work on modelling and analysis of gene regulation in cnidarians (corals and *Nematostella vectensis*), sponges, yeast and *Drosophila*. We do research on bio-mineralisation in corals and sponges (experimental and modelling work). We are working on growth and form of corals and the influence of light and hydrodynamics on the morphological plasticity and calcification in basal organisms (sponges and corals). This work is a combination of modelling work, a genetic comparison between different growth forms, phylogenetics, morphometrics of three-dimensional growth forms obtained from CT scans and experimental work.

Selected recent publications (3 max):

Fomekong Nanfack Y, Kaandorp JA & Blom JG (2007). Efficient parameter estimation for spatio-temporal models of pattern formation: Case study of *Drosophila melanogaster. Bioinformatics* 23, 3356-63.

Fomekong Nanfack Y, Postma M & Kaandorp JA **(**2009). Inferring *Drosophila* gap gene regulatory network: a parameter sensitivity and perturbation analysis. *BMC Syst Biol* 3: 94.

Tamulonis C, Postma M**,** Marlow H, Magie C, de Jong J & Kaandorp JA (2010). Morphometrics and Modeling of Gastrulation in the cnidarian *Nematostella vectensis Dev Biol* (in press)**.**

Key Personnel:

Dr. J.A. Kaandorp received his MSc in biology and a PhD in computer science and mathematics, both from the University of Amsterdam. Currently he has a permanent position as an associate professor at the Section Computational Science of the Faculty of Science of the University of Amsterdam. He runs a group of 2 MSc ,10 Phd students and 2 post-docs. The group is doing research at a range of different levels of organisation (genome-gene regulatory networks, cells- tissue-organism).

Dr. Carolina Cronemberger (post-doc) is a physicist by training and is currently working on modelling simulation of gene regulation, physiology and bio-mineralisation in cnidarians

Daniel Botman is trained as chemist and is doing his PhD on modelling of gene regulation of *Nematostella vectensis*

A post-doc to be hired will work on optimisation algorithms and modelling pattern formation in *Nematostella* development using high performance computing techniques

Partner 5: CWI

Description:

The Centrum Wiskunde & Informatica (CWI) is the Dutch national research institute for Mathematics and Computer Science. CWI is a private, non-profit organisation. Founded in 1946 (as Mathematisch Centrum), CWI aims at fostering mathematics and computer science research in The Netherlands. CWI receives a subsidy from the Netherlands Organization for Scientific Research NWO, amounting to about 70% of the institute's total income. The remaining 30% is obtained through national research programmes, international programmes and contract research commissioned by industry. CWI's mission is twofold: to perform frontier research in mathematics and computer science, and to transfer new knowledge in these fields to society in general and trade and industry in particular. The institute's strategy is currently inspired by four broad, societally relevant themes, a.o. Earth & Life Sciences.

CWI has always been very successful in securing a considerable participation in European research programs (ESPRIT, ACTS, TELEMATICS, BRITE, TMR, IST and others) and has extensive experience in managing these international collaborative research efforts.

Participating group: Scientific Computing for Systems Biology.

Role in the Project:

multi-scale modelling (ODE/DDE/PDE, stochastic CME/RDME/queuing theory models + verification (numerical analysis)); system identification (identifiability analysis, parameter estimation, global and local optimisation), model discrimination and optimal experimental design; resampling strategies/validation; uncertainty quantification; high performance computing (incl. GPU).

Expertise:

The group has strong expertise in scientific computing, in the last 8 years applied within systems biology. Emphasis lies in particular on: (i) multi-scale modelling: macroscopic ODE/DDE/PDE, mesoscopic CME/RDME/queueing theory models. Model assumptions, model building, implementation on various platforms and verification (numerical analysis). (ii) system identification: optimisation (local, global), optimisation measures, parameter estimation, model discrimination and optimal experimental design.

Selected recent publications (3 max)

Blom J & Mandies M (2011). Traffic generated by a semi-Markov additive process. *Prob Eng Inf Sci* 25: 1.

Ashyraliyev M, Fomekong-Nanfack Y, Kaandorp JA & Blom JG (2009). Systems bioloy: parameter estimation for biochemical models. *FEBS J* 276: 886–902.

Dobrzyński M, Vidal Rodriguez J, Kaandorp J & Blom J (2007). Computational methods for diffusion-influenced biochemical reactions. *Bioinformatics* 23:1969-77.

Key Personnel:

Joke Blom is a principal investigator and group coordinator at CWI in the Life Sciences group and is affiliated with the NISB (Netherlands Institute for Systems Biology). She is a mathematician by training. Her main research topic is scientific computing, in particular modelling (deterministic and probabilistic), numerical analysis, optimisation and model identification.

The postdoc to be hired will work on the (integration of the) modelling cycle with focus on system analysis (model validation and uncertainty quantification) and system identification.

Partner 6: FTELE.IGM

Description:

The Telethon Institute of Genetics and Medicine (FTELE.IGM) is an international reference centre for research on genetic diseases. It was created in 1994 by the Telethon Foundation, one of Italy's major non-profit organizations, to promote the advancement of research aimed at the diagnosis, prevention and cure of human genetic diseases. FTELE.IGM's mission is to understand the mechanisms of genetic diseases and to develop therapeutic and preventive strategies

Research activity at FTELE.IGM is supported by seven core facilities that provide state-of-the-art technology as well as "house-keeping" assistance. Each core is supervised by a FTELE.IGM investigator and is composed of specialized technical staff. Four cores (AAV vector Core, Microscopy and Imaging Core, Cell Culture and Cytogenetics Core, Transgenic and Knock-out Mouse Core Facility) offer high-quality and rapid scientific and technical services that help to improve and speed up the work of FTELE.IGM investigators. The Bioinformatics Core offers expertise in exploration and analysis of experimental data (statistical data analysis, sequence data analysis) to help investigators in the Institute with their research. Finally, the Informatics Core and the General Services Core provide maintenance for the Institute's general activities and resources.

Role in the Project:

Development of algorithms to reverse-engineer gene regulatory networks from gene expression data and to identify drug mode of action.

Expertise:

FTELE.IGM is expert in reverse-engineering gene regulatory networks from high-throughput data both in yeast and mammalian cells using differential equations and information theoretic approaches. In addition, FTELE.IGM is also expert in Synthetic Biology specifically in the construction and modelling of synthetic regulatory circuits in yeast and mammalian cells.

Selected recent publications (3 max):

Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, Murino L, Tagliaferri R, Brunetti-Pierri N, Isacchi A & di Bernardo D (2010). Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc Natl Acad Sci USA* 107:14621–6.

Cantone I, Marucci L, Iorio F, Ricci M, Belcastro V, Bansal M, Santini S, di Bernardo M, di Bernardo D & Cosma MP (2009). A Yeast Synthetic Network for In Vivo Assessment of Reverse-Engineering and Modeling Approaches. Cell 137: 172–81.

Bansal M, Belcastro V, Ambesi-Impiombato A & di Bernardo D (2007). How to infer gene networks from expression profiles. *Mol Syst Biol* 3:78.

Key Personnel:

The research team will be led by myself (Diego di Bernardo) directing and supervising post-doctoral fellows and graduate students, performing experiments, making decisions on research strategies, writing publications and presenting results at scientific conferences. We are requesting funding to cover a graduate student with a background in Computer Science or Engineering.

Partner 7: UNIMAN

Description:

The University of Manchester (UNIMAN) is the largest university in the UK. The team participating in this project belong to the Manchester Centre of Integrative Systems Biology (MCISB), one of the six BBSRC-funded national centres for systems biology. UNIMAN and MCISB have extensive experience in both participating and managing research projects and training grants, at the national as well as European level. MCISB is located in the Manchester Interdisciplinary Biocentre which hosts academics from a wide range of disciplines from 3 different Faculties.

Role in the Project:

Large-scale modelling; multi-scale modelling; global sensitivity analysis; software development.

Expertise:

The MCISB has strong expertise in all aspects of systems biology, the Mendes group within it has special emphasis on:

(i) development of software infrastructure and standards for systems biology (COPASI, SBML, SBRML, several data and model management packages),
(ii) modelling and simulation of large scale metabolic networks
(iii) global sensitivity analysis
(iv) enzyme kinetics for systems biology

Selected recent publications (3 max):

Smallbone K, Simeonidis E, Swainston N & Mendes P (2010). Towards a genome-scale kinetic model of cellular metabolism. *BMC Syst Biol* 4: 6.

Dada JO, Spasic I, Paton NW & Mendes P (2010). SBRML: a markup language for associating systems biology data with models. *Bioinformatics* 26: 932–8.

Sahle S, Mendes P, Hoops S & Kummer U (2008). A new strategy for assessing sensitivities in biochemical models. *Phil Trans Roy Soc A* 366: 3619–31.

Key Personnel:

Prof. Pedro Mendes is the Chair in Computational Systems Biology in the School of Computer Science and the Deputy Director of the Manchester Centre for Integrative Systems Biology. Mendes is also a Research Professor in the Virginia Bioinformatics Institute at Virginia Tech (20% appointment). He obtained his PhD in Biochemistry from the University of Wales Aberywstwyth in 1994, where he also was a post-doc until the end of 1998. From 1999-2000 he was the Program Leader for Pathways at the National Center for Genome Resources (Santa Fe, NM, USA), from 2000 onwards he has been a Professor at the Virginia Bioinformatics Institute (Assistant Prof, Associate Prof and now Full Prof) but from 2007 onwards he reduced his appointment there to 20%, taking up a Chair position in the University of Manchester (80%). He is also an Adjunct Professor at the Wake Forest University Medical School. Currently he is a member of the Editorial Board of *IET Systems Biology* and *Transactions on Computational Systems Biology*. He is a member of the BBSRC Committee C for grant reviews. Mendes has published over 70 publications, with a H-index of 25, and average 54 citations per paper.

Mendes currently leads a group of 7 Research Associates and 12 PhD students. The research activities of his group are: 1) development of the widely used biochemical simulator COPASI and other software applications for systems biology, 2) enzyme kinetics characterization for systems biology models, 3) construction of large scale metabolic models (flux balance analysis and full kinetic models), 4) multi-scale modelling, 5) data mining in systems biology, and 6) modelling specific biochemical processes (interleukin-1 signalling, iron metabolism, the pentose phosphate pathway, hepatitis C infection, growth and division in yeast).

Partner 8: USheff

Description:

The University of Sheffield recently appointed Neil Lawrence and Magnus Rattray to cross faculty positions to take a leadership role in computational systems biology and bioinformatics. Lawrence and Rattray are part of a new centre for mathematical modelling of biological systems which draws members from across the University. Their research interest is in the integration of mathematical models with biological data to reverse engineer the fundamental interactions within biological cells.

Role in the Project:

Data integration and visualisation, parameter estimation, dealing with parameter sloppiness, model ranking and applications in signalling cascades.

Expertise:

Lawrence and Rattray are acknowledged experts on integration of mechanistic models, based around differential equations, with probabilistic approaches to allow for a rigorous Bayesian analysis of a biological system. These approaches are particularly appropriate for computational systems biology where the data is typically sampled more sparsely and with higher noise than in traditional engineering systems. Their background is the statistical machine learning community and their expertise extends to latent variable modelling including non-linear probabilistic latent variable models.

Selected recent publications (3 max):

Honkela A, Girardot C, Gustafson EH, Liu YH, Furlong EEM, Lawrence ND & Rattray M (2010). Model-based method for transcription factor target identification with limited data. *Proc Natl Acad Sci USA* 107: 7793–8.

Pearson RD, Liu X, Sanguinetti G, Milo M, Lawrence ND & Rattray M (2009). Puma: a Bioconductor package for propagating uncertainty in microarray analysis. *BMC Bioinformatics* 10: 211.

Lawrence ND (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *J Mach Learn Res* 6: 1783–1816.

Key Personnel:

Professor Neil Lawrence has a background in machine learning and computer science. After an undergraduate degree in Mechanical Engineering he completed his PhD with Professor Chris Bishop at the Computer Lab in Cambridge. His main expertise is probabilistic modelling with applications. He has considered applications to data such as speech, vision and robotics, but his main application focus is computational biology, with a particular interest in reverse engineering biological systems through probabilistic modelling.

Professor Magnus Rattray also has a background in machine learning and computer science. He completed his undergraduate degree in Physics before studying for a PhD using statistical physics for the analysis of genetic algorithms. Since 1998 he has focussed on applications of machine learning and statistical methodologies in biological applications, including phylogenetics and reverse engineering of biological systems.

A post-doc to be hired on this project will be developing new probabilistic methodologies for integrating biological data with mechanistic models for computation of Bayes factors and with data visualisation algorithms based around probabilistic latent variable models.

Partner 9: CSM

Description:

The CoSMo Company (incorporation in June 2010) is a very young spin-off company of both the CNRS (the French National Agency of Research) and Ecole Normale Supérieure de Lyon (one of the top scientific “Grandes Ecoles”). The goal of CoSMo is to conceive, implement, and disseminate state-of-the-art software tools dedicated to in silico concrete problem solving for complex systems. In order to facilitate a broad (academia and industry) acceptance of its modelling and simulation platform, CoSMo adopts an open-access software license (BSD) and as such is incubated at IXXI one of the leading French complex systems institutes. Founders of the company co-ordinated and participated (as WPL) to several European (FP6 and FP7) as well as national scale (ANR) research projects on various biological scientific domains like embryogenesis (animals), morphogenesis (plants) as well as epidemiology. As a company CoSMo is actively involved in research collaboration agreements in the field of immunology (Singapore Immunology Network) as well as one of the top 10 French industrial companies.

Role in the Project:

To integrate—within an open modelling and simulation platform—the required tools for the full model-building cycle starting from model implementation (multi-scale, trans-scale, hybrid, possibly geometrical models), to their reconstruction, visualisation, study, and hopefully their validation (with a simulation approach) against integrated databases.

Expertise:

CoSMo has expertise in integrated yet open software tools dedicated to model integration of multi-scale biological systems, with emphasis on:

- portable modelling languages (including dynamical aspects) for systems biology;
- integration of sub-models including over various time scales;
- spatial models;
- visualisation of dynamics of large scale networks;
- model study with a strong numerical simulation approach.

Selected recent publications:

Not applicable. As technology-oriented researchers, we publish our code.

Key Personnel:

Presently, the research activities of CoSMo are carried out by 8 persons:

- Eric Boix, CSO, who was Work Package Leader in charge of the modeling and simulation platform within the European project Morphex (FP6, http://morphex.org) and Dynanets (FP7, http://dynanets.org), received his PhD in mathematics in 1994 while studying formal discrete equivalent of geometrical invariants with an initial computational approach, and integrated the computer science laboratory at ENS Lyon where he took part to the development of middle-ware grid computing software (DIET) and mainly to the complex system simulation group within IXXI (Complex Systems Institute).
- Michel Morvan, scientific advisor, was coordinator of the European project Morphex, professor of computer science at ENS Lyon, research director (“directeur d'études”) at the Ecole des Hautes Etudes en Sciences Sociales (School of High Studies in Social Sciences) in Paris and former member of the Institut Universitaire de France. Since July 2004, he is External Faculty of the Santa Fe Institute. His research took originally place in the context of theoretical computer science and discrete mathematics, by the end of the 90s, he has oriented his research in the direction of complex systems and created in Lyon the "Institut des Systemes Complexes - Complex Systems Institute".
- 6 permanent software engineers which high standards for scientific software.

Partner 10: INSIL

Description:

Insilico Biotechnology is a privately owned company located in Stuttgart, Germany. It designs and optimises biotechnological processes for the chemical and pharmaceutical industries. Successful in business since 2001, Insilico has internationally renowned expertise and a unique technology platform for connecting cell model libraries with simulation processes. Insilico analyses the latest biotech data and integrates it in genome-wide network models. With its high-performance computing techniques, Insilico develops new improved solutions for manufacturing biochemicals and biopharmaceuticals and achieves considerable cuts in the time needed for the development of bioprocesses.

Role in the Project:

Provision of large-scale dynamic networks, high-performance computing

Expertise:

Insilico participated in a number of relevant project including HEPATOSYS (BMBF, Germany): systems oriented analysis of detoxification in hepatocytes; ZIM-HPC (BMWi, Germany): application of high performance grid computing for identifying systems dynamics in large-scale networks; and MedSys (BMBF, Germany): A systems oriented approach to cell-tissue interaction. Insilico partners in the FP7 Alternative-Testing-Strategies (Cosmos and Notox) as well as in the Virtual Liver Network (BMBF, Germany). Key expertise of Insilico Biotechnology:

- Graphically oriented reconstruction of genome-based networks
- Parameter estimation and network verification
- Simulation and analysis of intracellular fluxes
- High-performance computing
- Data integration

Selected recent publications:

Maier K, Hofmann U, Reuss M & Mauch K (2010). Dynamics and Control of the Central Carbon Metabolism in Hepatoma Cells. *BMC Syst Biol* 4:54.

Maier K, Hofmann U, Bauer A, Niebel A, Vacun G, Reuss M & Mauch K (2009). Quantification of statin effects on hepatic cholesterol synthesis by transient (13)C-flux analysis. *Metab Eng* 11: 292–309.

Maier K, Hofmann U, Reuss M & Mauch, K (2008). Identification of Metabolic Fluxes in Hepatic Cells from transient (13)C Labeling Experiments: Part II Flux Estimation. *Biotechnol Bioeng* 100: 355–70.

Key Personnel:

Dr. Joachim Schmid is group leader of the Industrial Biotechnology group. He received his Master degree in Chemical Engineering. After a PhD on a systems-oriented approach to *E. coli* metabolism from the University of Stuttgart, he joined Insilico.

Dr. Dirk Müller is leader of the Biopharma group at Insilico Biotechnology. Before joining Insilico, Dirk Müller was a Post-doctoral Fellow with Prof. J. Stelling at the Institute of Computational Science (ETH Zürich) focussing on signal transduction and gene regulation in yeast.

Dipl.-Inform. Anne Bonin received her Diploma degree (M.Sc.) in Bioinformatics from the University of Tübingen. At Insilico Biotechnology, she is project leader for High Performance Grid Computing and has led several research projects in the area of inferring large-scale network dynamics.

Partner 11: FS

Description:

FS is a global supplier of ingredients with beneficial health effects. The company develops processes for the production of nutraceutical ingredients by fermentation of metabolically engineered microorganisms, produces the ingredient at toll manufacturers, has its own sales force and supplies the ingredient as raw material particularly to the dietary supplement and food industry, but also to cosmetics companies. Today, FS has just released two products on the market, trans-resveratrol and 1,3-1,6 beta-glucan. The former product has been developed by FS. The latter product has been taken into the product portfolio by FS by gaining exclusive rights from GlycaNova to market and sell the product in the US dietary supplement market. The company has currently a team of 16 researchers (30 employees in total) that conduct metabolic engineering, fermentation and analysis. FS is currently aiming at developing the production of the omega-3 and omega-6 fatty acids which is targeted to be the next product in FS' product portfolio, and further nutraceutical ingredients.

Role in the Project:

Software tester (preparation of recommendations for software); modelling of metabolism of nutraceutical ingredient producing micro-organisms, particularly *Saccharomyces cerevisiae*; simulation of nutraceutical ingredient production; generation of experimental data (flux, transcription and metabolite-level data) for model validation and improvement of models.

Expertise:

FS uses a strong metabolic engineering and synthetic biology platform for the design of nutraceutical ingredient producing micro-organisms. This includes extensive expertise in:

(i) genetic engineering (incl. protein engineering) of microorganisms,
(ii) fermentation (batch, chemostat, fed-batch) at lab-scale (300 ml – 5 L),
(iii) analysis of intracellular and extracellular metabolites,
(iv) scale-up of fermentation and down-stream processing processes including production,
(v) modelling of metabolism of microorganims, particularly *S.cerevisiae.*

Selected recent publications (3 max):

WO2005118814 (Patent). Metabolically engineered cells for the production of polyunsaturated fatty acids.

WO2008000277 (Patent). Microbial bioreaction process.

Tavares S, Grotkjær T, Obsen T, Haslam RP, Napier JA & Gunnarsson N (2010). Metabolic engineering of *Saccharomyces cerevisiae* for production of eicosapentaenoic acid using a novel D5-desaturase from *Paramecium tetraurelia*. *Appl Environ Microbiol* (in press).

Key Personnel:

Dr. Jochen Förster: COO and co-founder of FS. Jochen has Ph.D. in Biotechnology from the Technical University of Denmark and is part of the management of FS with the responsibility to overlook the project portfolio of FS, which involves the development of novel biotech processes for production of nutraceutical ingredients. He is inventor of 8 patent applications and co-author of > 15 scientific publications. He has extensive management experiences, and he also has solid experience with EU projects from previous and current projects including Combig-Top, YSBN, where he acted as WP leader, and SysInBio.

Dr. Hans Peter Smits: Head of Fermentation Department, Hans Peter holds a Master degree in Biology /Biochemistry from Utrecht University in The Netherlands and a Ph.D. in Chemistry from Amsterdam University in The Netherlands. Before joining Fluxome, Hans Peter was Assistant Professor at the Technical University of Denmark. Dr. Hans Peter Smits is co-author of more than 10 scientific publications and is inventor of 5 patent applications.

To be identified FS intends to employ one bioinformatician with strong background in physiology and modelling of metabolism.

2.3 Consortium as a Whole

The problems targeted in this project cover a wide range of scientific disciplines and scientific/technological applications. Therefore, they require a multi-disciplinary, community-based approach, since they cannot be solved by any single research group. The BioPreDyn consortium brings together the necessary range of overlapping, but complementary backgrounds and competences to ensure a successful project. The consortium includes top European academic groups, plus three SMEs from seven European countries with synergistic expertise in areas including databases, scientific visualisation methods, statistics, machine learning, mathematical modelling, and biotechnological (bio-process) engineering.

The consortium combines geographical and disciplinary diversity with academic and biotechnological excellence. The partners complement each other in useful and synergistic ways. For example, in the case of mathematical modelling of biological systems, members of the consortium literally cover the entire modern spectrum of techniques, from data analysis and visualisation, to machine learning and data-driven dynamical modelling, to global non-linear optimisation, to model and parameter analysis, model discrimination and optimal experimental design.

The three SMEs participate in the project in order to adequately implement and exploit the results of the project. These companies provide complementary and suitable expertise and applications for the objectives of BioPreDyn: a life sciences software company (CoSMo, CSM), an industrial biotechnology company (Fluxome, FS) and a bioprocess engineering company (Insilico Biotechnology, INSIL).

The collective expertise of the active partners is reinforced by a world-class Scientific Advisory Board (SAB), which includes Prof. Hiroaki Kitano (Sony Computer Science Laboratories, Japan), Prof. Francis Doyle III (UC Santa Barbara, USA), Dr. Nicolas Le Novère (EBI-EMBL, UK), Prof. Roel van Driel (SILS/Univ. of Amsterdam, NL) and Prof. Victor de Lorenzo (CNB-CSIC, ES).

Most of the partners have previously collaborated with other participants in the consortium (see below). They are fully committed to the project, and have ample expertise and experience in the fields of activities covered by it, offering the capacity and resources to fulfil the project objectives. The suitability and commitment of each academic partner, together with their current collaborations, are detailed as follows:

Partner 1: CRG

- *Suitability:* Our group specialises in reverse-engineering developmental gene regu-latory networks based on spatial time series of quantitative expression data. Our experience lies in the application of cutting-edge modelling and optimisation algorithms to complex biological systems.
- *Commitment:* The principal investigator, two post-docs, and a programmer/computer technician.
- *Existing links with other partners:* (1) Collaboration (funded by the ComplexityNET scheme) with UvA on multi-objective, non-linear, global optimization. (2) Informal collaboration with CWI on parameter estimation and identifiability analysis. (3) Informal collaboration with USheff on inference of missing state variables and reverse-engineering of Drosophila mutants. (4) Informal collaboration with CSIC on global optimisation (scatter search, meta-heuristics).

Partner 2: CSIC

- *Suitability:* Our group has strong expertise in dynamical modelling and optimization of biological systems, with emphasis on robust parameter estimation, optimal experimental design and identifiability analysis and optimal control of biosystems.
- *Commitment:* Three senior researchers, five post-docs and several PhD students.
- *Existing links with other partners:* CSIC has collaborated with UNIMAN on parameter estimation in systems biology, and is currently having a similar collaboration with CRG. CSIC is also collaborating with EMBL in optimization in computational systems biology.

Partner 3: EMBL

- *Suitability:* Our group has expertise on multidimensional data processing, and visualisation. Mathematical modelling of signalling networks with different mathematical formalisms, with focus on large networks integrated with high-throughput data.
- *Commitment:* One senior researcher, one scientific programmer, one post-doc.
- *Existing links with other partners:* EMBL collaborates with CSIC on parameter estimation and other optimization problems, and with FTELE.IGM on data analysis/visualisation and network modelling.

Partner 4: UvA

- *Suitability:* The UvA (Section Computational Science) has long-standing experience in high performance computing, scientific visualization and modelling and simulation in computational biology, bio-medical applications and physics. Within computational biology we do research at a range of different levels of organisation (genome-gene regulatory networks-cells-tissue-organism).
- *Commitment:* One senior researcher, one post-doc and several Phd students.
- *Existing links with other partners:* The UvA (Section Computational Science) has a collaboration with the CRG in computational systems biology and has a long-standing collaboration with CWI in several systems biology projects.

Partner 5: CWI

- *Suitability:* Our group has expertise in mathematical and computational modelling of biochemical systems including system identification/optimal experimental design.
- *Commitment:* One senior researcher, one post-doc.
- *Existing links with other partners:* CWI has collaborated with the CRG on model identification and with UvA on modelling (deterministic/stochastic) and parameter estimation.

Partner 6: FTELE.IGM

- *Suitability:* This group has a strong expertise in reverse-engineering of gene regulatory networks from gene expression data and building quantitative mode of gene regulation. In addition new Systems Biology approaches to identification of drug mode of action has been developed.
- *Commitment:* One senior researcher, one PhD student and one post-doc.
- *Existing links with other partners:* FTELE.IGM is collaborating with EMBL.

Partner 7: UNIMAN

- *Suitability:* This group is a pioneer in application of optimization algorithms in biochemical modelling and is one of the authors of the widely used software for systems biology simulation (COPASI), and are active participants in the SBML community effort. We have also a strong track record in reconstruction of metabolic networks and generating large-scale kinetic models from these. The group is an active member of the Manchester Centre for Integrative Systems Biology, which is establishing methodologies for bottom-up systems biology, particularly in *S. cerevisiae*. The Mendes group has also published research in reverse-engineering gene networks
- *Commitment:* The principal investigator, one post-doc, and one PhD student.
- *Existing links with other partners:* UNIMAN has collaborated in parameter estimation algorithms with CSIC, and in establishing standards for reverse-engineering with FTELE.IGM.

Partner 8: USheff

- *Suitability:* This group has strong expertise in probabilistic modelling applied to systems biology problems such as: Bayesian parameter estimation, model-based ranking of transcription factor targets and regulatory network inference from time-series data, Bayesian model selection and experimental design, High-throughput genomic and epigenomic data processing. The group also develops novel non-linear dimensionality reduction and visualisation techniques.
- *Commitment:* Two principal investigators, one post-doc.
- *Existing links with other partners:* USheff has an informal collaboration with CRG on inference of missing state variables and reverse-engineering of *Drosophila* mutants.

Industrial/Commercial Involvement: Participation of SMEs

The consortium incorporates three high-profile companies with different yet complementary profiles:

- *Complex Systems Modelling/CoSMo (CSM)* is a software company specialised in complex systems modelling and simulation with a focus on systems biology. CSM expects to benefit from the wide range of biological problems addressed by the project as well as the diversity of the numerical methods, which will help CSM consolidate its know-how and expertise concerning the systems biology modelling cycle.
- *Insilico Biotechnology (INSIL)* designs and optimises biotechnological processes for the chemical and pharmaceutical industries. This SME can benefit from the development of novel model-building strategies and their application to large-scale kinetic models of microorganisms integrated with regulatory and signalling networks.
- *Fluxome SA (FS)* is an industrial biotechnology company, which develops processes for the production of nutraceuticals (ingredients with beneficial health effects) by fermentation of metabolically engineered microorganisms. Fluxome can greatly benefit from model-based methods to be developed in this project in order to optimise their processes and to guide the metabolic engineering procedures.

These SMEs have the following plans to ensure the exploitation of results:

- *CSM* plans to disseminate (freely for academics, commercially for corporate clients) the integrated software framework containing the numerical tools to be developed in this project, as a contribution to strengthen the European systems biology software community. It is CoSMo's direct interest to ensure this form of exploitation since dissemination of this code framework will illustrate CoSMo's expertise in scientific software development.
- *INSIL* offers high-tech solutions and services to the Life Science industries and expects the BioPreDyn project to greatly increase its competitiveness and give it an edge over competitors, in particular from North America and Asia. Project results will enable *INSIL* to provide customers with novel solutions adding value at different points along the value chain in the future. New tools and methods developed within this project will significantly accelerate both model development and model verification. In combination with the envisaged integrated network models, this is a key prerequisite for entering new application areas in industrial biotechnology and in the manufacturing of biopharmaceuticals. These application areas include the prediction of gene targets for improving the product yield for fine chemicals such as succinic acid, methionine, and vitamine B2 using microbial strains like *E. coli*. These predictions can then capitalize on network models combining the interaction of metabolism, gene regulation and/or signalling processes. For the production of therapeutic antibodies using CHO cell cultures, large-scale dynamic models will pave the way for predicting the impact of relevant process variables like pH and/or media composition on cell growth and productivity or regarding clinically important aspects of product quality, such as glycosylation patterns. Such

predictions are notoriously difficult or unreliable using today's methods. Through collaborations within the consortium, *INSIL* will gain access to new know-how, which it wants to exploit for extending its range of services offered. During the project, *INSIL* plans to publicly advertise the BioPreDyn project through announcements on the company homepage and inclusion in its marketing materials such as flyers and customer presentations. *INSIL* is going to disseminate project results on conferences and will use these for acquiring new customers and partners at the end of the project.

- *FS:* Results from BioPreDyn will be used in FS's ongoing development projects that focus on the production on resveratrol and PUFA production. The models will increase the understanding of heterologous biosynthesis of such nutraceutical ingredients in baker's yeast. All models that will be built within this project are planned to become an integral part of the technology platform of FS. Hence, this will strengthen and extend particularly FS modelling platform and will find application in the design of improved and other novel bioprocesses. Altogether, the tools and models of BioPreDyn have clearly the potential to decreasing the time to market of novel products. In order to secure application of such models beyond BioPreDyn, FS aims at employing further FTEs beyond the BioPreDyn Project lifetime. Furthermore BioPreDyn will lead to the identification of novel metabolic engineering and synthetic biology strategies. Such strategies will be tested experimentally outside the BioPreDyn project. In the cases of confirmation of modelling results by wet lab experiment, patent protection of the most promising strategy is planned, according to the BioPreDyn Consortium Agreement.

2.4 Resources to be Committed

The following table illustrates the total budget and EC contribution breakdown among the partners, according to cost categories and activities:

Particpants	RTD Personnel	RTD Consumables	RTD Equipment	RTD Travel	RTD Other	RTD Overheads	Management	Other activities	TOTAL
1-CRG	200.000	0	4.000	9.000	5.000	130.800	184.388	32.000	565.188
2-CSIC	138.450	5.000	0	9.000	0	274.410	0	0	426.860
3-EMBL	135.610	2.500	0	4.500	0	85.566	0	0	228.176
4-UvA	202.301	5.000	0	25.000	0	177.356	0	0	409.657
5-CWI	189.141	0	1.500	6.000	3.000	156.930	0	0	356.571
6-FTELE.IGM	120.000	0	0	6.000	0	25.200	2.000	0	153.200
7-UNIMAN	183.879	6.390	2.556	21.303	0	128.477	1.500	0	344.105
8-Usheff	230.030	3.000	2.000	6.000	0	144.618	2.000	0	387.648
9-CSM	225.000	3.000	0	7.000	0	141.000	0	0	376.000
10-INSIL	180.000	30.000	0	6.000	0	129.600	0	0	345.600
11-FS	199.145	3.000	0	6.000	0	124.887	0	0	333.032
Total eligible	2.003.556	57.890	10.056	105.803	8.000	1.518.844	189.888	32.000	3.926.037
EC contribution	1.502.667	43.418	7.542	79.352	6.000	1.139.133	189.888	32.000	3.000.000

Table 1: BioPreDyn budget and EC contribution breakdown.

RTD Activities (92,6% of total EC contribution)

Personnel – The envisaged efforts to accomplish the project goals will encompass 419,9 person-months for RTD activities over 36 months (50,1% of RTD EC contribution).

Consumables and Equipment – The laboratories at all partner sites are extremely well equipped to conduct the research proposed, as described below. As a result, the major resources required for BioPreDyn, in addition to personnel, are workstations and software, as well as contributions to the maintenance and running costs of computer clusters, which are included under consumables or equipment, depending on the normal institution practices. These costs represent 1,7% of the RTD EC contribution. Specifically, partner 10 (INSIL) is budgeting a contribution to maintenance and running costs of their high-performance computer cluster (Intel Nehalem architecture with 5,600 cores).

Travel – The travel budget (2,6% of RTD EC contribution) will be used for project meetings, visits to partner sites and participation in scientific conferences.

Other – This category (0,2% of RTD EC contribution) covers other costs such as publication costs and conferences fees.

Overheads – Most partners use the special transitional flat rate (60%), with the exception of partner 2, CSIC, which uses the real indirect cost method with an overhead of 180%. This represents 38% of the RTD EC contribution.

Management Activities (6,3% of total EC contribution)

The management budget will mainly include personnel costs for a part-time project manager (€90.000, Partner 1, CRG), project meetings and the participation of Scientific Advisory Board members (€14.742, Partner 1, CRG), gender budget to be assigned (€6.000), a laptop for the project manager (€2.000), and audit certificates for the partners requiring them (€9.500).

Other Activities (1,1% of total EC contribution)

Other activities cover **dissemination** (website, leaflets, posters, and stalls at conferences/trade fairs) and **training** (two workshops organized by two different partners (€20.000 to be assigned by the coordinator). Publications in peer-reviewed journals and presentations at conferences will be covered by the partners' budgets.

Additional Resources of the Participants

Apart from the additional costs budgeted in section A3.1 as total budget, all partners already dispose of cutting-edge equipment and infrastructure at their laboratories (e.g. networked high-end workstations with the required software for code development and testing, and/or servers for hosting of databases), and permanent staff, which they will provide and use for project work without charging related costs to BioPreDyn:

Participant	Infrastructures	Additional Personnel
CRG	Access to the Mare Nostrum super-computer (>10'000 cores connected by Myrinet), run by the Barcelona Supercomputing Center (BSC; www.bsc.es), granted on a three-month, project-specific basis; as well as access to a CRG in-house cluster (~200 cores), mainly for testing and calibrating software.	Dr. Anton Crombach (post-doc) Damjan Cicin-Sain (Programmer)
CSIC	CSIC's HPC cluster with 98 cores and access to the HPC facilities at CESGA (www.cesga.es), which includes the Finisterrae super-computer, currently the third most powerful in Spain.	Dr Antonio A. Alonso, Dr. Balsa-Canto
EMBL	EBI hosts a number of databases, including ArrayExpress (gene expression), Ensembl (Genomics), PRIDE (proteomics), and IntAct (proteininteraction networks), which will be used extensively during this project.	Jerry Wu (scientific programmer)
UvA	The Section Computational Science has access to several large-scale facilities (e.g. the Lisa computing cluster in Amsterdam, the DAS-II distributed computing cluster in the Netherlands) and has a fully-equipped visualization lab.	Dr. Carolina Cronemberger (post-doc)
CWI	CWI has an excellent IT environment, it owns (among others) a Linux cluster (48 64-bit dual Opterons), and the group has access to the HPC systems of SARA (subtrac.sara.nl/userdoc), the Dutch National High Performance Computing and e-Science Support Center, and the Dutch supernode in the International Science Grid.	none
FTELE.IGM	FTELE.IGM's Bioinformatics and Informatics cores will provide support for the installation and maintenance of our relational database infrastructure.	Post-doc (to be hired)
UNIMAN	The UNIMAN group has access to the computational resources of the Manchester Centre for Integrative Systems Biology, composed of two 16-core servers, which are dedicated to website and database servers, and access to a large CONDOR pool of over 1500 cores that are available for high-performance computing in the Faculty of Engineering and Physical Sciences of the University of Manchester.	PhD student (to be hired)
USheff	Our group has access to the computational resources of the University of Sheffield.	none
CSM	CoSMo shall provide the servers for hosting the website, the collaborative code development framework (SVN, Trac) as well as the agile programming platform (Cdash) for continuous software builds in order to assert code quality.	6 software engineers (providing support)
INSIL	In-house high performance computer cluster (Intel Nehalem cluster with 5,600 cores).	Dr. Dirk Müller Anne Bonin
FS	Fluxome has the required lab infrastructure to collect additional experimental data on their *S. cervisiae* production strains if required.	Dr. Hans Peter Smits

3. Impact

3.1 Expected Impacts Listed in the Work Programme

The BioPreDyn project aims at developing new bioinformatics methods and tools for data-driven and predictive dynamic modelling with the final goal to better understand specific biological questions and datasets, as well as to implement and pave the way to new biotechnological applications. By bridging multiple disciplines (from bioinformatics, systems biology, microbiology to biotechnology), and interlinking diverse players (universities, research centres, international organizations and SMEs) the project will have a profound impact (as described more extensively in the following sections), matching the expectations of the call, the KBBE Work Programme 2011, and the Europe 2020 strategy.

Better Exploitation of Existing Databases

'Omics' tools, the high-throughput methods to characterize genes, proteins, small molecules and their interactions in a precise, quantitative and dynamic fashion, are continuously being improved and applied to a wide range of complex systems in biology. This trend poses the great challenge of making sense out of the enormous amount of data we are producing. BioPreDyn will strive to better exploit existing databases by developing tools, methods and workflows for semi-automated **data integration and visualisation** (WP1/2). Moreover, the consortium will find solutions to guide the user in dealing with dynamic expression data, with data across space, incomplete, heterogeneous or noisy data. Within WP3 (and also WP8), software tools and workflows will be integrated in a **single computational framework** that will support the entire **systems biology modelling cycle** (Fig. 1) overcoming many current problems in modelling: software often too difficult to use, software not compatible or interoperable, different languages and data formats. Finally, these concerted efforts will lead to an increased predictive and interpretative capacity of the available data.

The infrastructure that BioPreDyn generates will be made available to the scientific community (freely) and to the private sector: datasets will be integrated in the NetBase infrastructure provided by FTELE.IGM, while CSM will provide a centralized, and standardised software suite with graphical interfaces for our tools. Thanks also to our dissemination and training activities, this will have a profound impact not only on research in general, but also on the translation of research findings and new methodology into new biotechnological (and medical) applications.

Paving the Way for New and Optimised Biotechnological Applications

The computational tools, software and workflows generated by BioPreDyn will be of general use. As proof of concept, partners will apply this infrastructure to a selected set of fundamental biological questions and biotechnological applications. This parallel way of operating will also facilitate the dialogue and **knowledge sharing between the academic and the industrial partners**, since most likely they will face similar technical and methodological problems. In addition, new findings in basic research (WP4–6) will pave the way to novel strategies in development of biotechnological processes (WP7).

Despite the fact that microorganisms can play harmful roles, they are nowadays becoming crucial in solving a wide range of societal, environmental, health, and economical problems. Microorganisms can in fact be engineered to provide alternative sources of energy or bulk chemicals, clean up the environment from wastes and toxic substances, produce food additives (nutraceuticals) or therapeutic proteins, as few examples. We have at our disposal a vast biological knowledge and battery of tools to modify microorganisms, but we have learned in the past that microbial engineering requires a system-wide approach rather than a reductionist way by intervening on a single gene or protein. Recent news such as the engineering of *E. coli* to produce different types of biodegradable plastic and the getting closer to the production stage of bacterial fuel production show that the field offers exciting opportunities that BioPreDyn does not want to miss.

WP4 will focus mostly on large-scale models of microbial and eukaryotic cells. FS uses as preferred microorganism *S. cerevisiae* for the production of nutraceuticals, such as PUFA (long chain omega 3), traditionally recovered from fish stocks. Taking into account that fish stocks are

today under extreme pressure worldwide—and some studies even predict the collapse of commercially exploited fish stocks—*S. cerevisiae* large-scale dynamic modelling (WP4) will likely provide new insights in microbial physiology to implement fast, safe, efficient and cost-effective production of PUFA from sustainable natural sources (WP7). INSIL will benefit of basic findings to optimise new pathways**,** productivity and specific characteristics of bacterial strains for biotechnology-based production of a diverse range of biopharmaceuticals and fine chemicals.

Benefit for Academia and SMEs

BioPreDyn will generate a mutually beneficial partnership between eight academic labs and three SMEs. Academic groups will benefit from the establishment of proof-of-concepts for the technical and economical exploitation of the know-how and infrastructure generated by the project. The SMEs will highly benefit from this collaboration since they will broaden their product and technology portfolio, such as software and computational tools (CSM), and they will profit from model-based optimisation for their modelling tools (INSIL), and production bio-processes (FS). Overall, synergies between academic partners and SMEs catalysed by BioPreDyn will facilitate the development and application of microorganisms in industrial and medical biotechnology, and contribute to shortening time to market (from idea to market). In general, competiveness of SMEs will be strengthened and properly equipped to take on competition not only from the US but also from emerging countries such as China, India and Brazil.

Boosting European Innovation

Research and innovation (Europe 2020) are at the core of BioPreDyn. The project will pursue a holistic approach from basic research to translating the project results to developing markets in biotechnology, making a step forward in **bridging the so-called innovation gap**.

Since the three SME partners work in close collaboration with other industries, the beneficial impact of BioPreDyn on biotechnology will be further amplified. FS, as example, develops production processes for nutraceutical ingredients, and delivers its protocols to larger companies that implement production of nutraceuticals or similar ingredients on a large scale. FS sells its products to the dietary supplement industry, an industry that more and more demands products at constant high quality and at low price. INSIL predicts and optimizes microbial biotechnological processes for the food, agro, and healthcare industries, collaborating with major players in the field, such as Bayer Technology Services, Boehringer Ingelheim, and DSM Food Specialty.

Finally, the project will **integrate education and innovation aspects** through inter-sectorial (academic labs and SMEs) visits of young researchers working in the project, one specific workshop during the first year, and active participation, monitoring and mentoring by the Innovation Board.

3.2 Dissemination/Exploitation of Project Results, and IP Management

Training, dissemination and exploitation activities are central to the project and dedicated work packages (WP8 & WP9) have been designed for their implementation at the highest standards. The following sections describe the strategy for each activity in more detail.

3.2.1 Training

The interdisciplinary nature of the project will offer great training opportunities for the junior researchers involved (PhD students and post-doctoral research fellows). Moreover, mobility of researchers will be promoted among labs and among academic groups and SMEs (see Table 1.3d, WP9). More specifically, the following training activities will be organized:

- **Two week-long workshops** (CRG and EBI/EMBL). The 1st workshop will be open only to people directly involved in the project, and the 2nd will aim at training scientists outside the consortium in the methods developed during this project. Both workshops will focus on the state-of-the-art and novel methods and computational tools to better exploit databases, integrate and visualize data, and build and validate computational models. International experts will be invited to contribute to both workshops. The 1st workshop will also offer training on intellectual property, technology transfer, and innovation under supervision of the Innovation Board. Both workshops will also include a session to illustrate case studies where

our tools will be applied to specific and applied problems in biotechnology, including a discussion on ethics and their impact on society at large.

- **Short-term and medium-term exchange scheme**. Partners will encourage exchange visits among the labs, especially between the academic and private sectors. This scheme will facilitate knowledge transfer and open wider career opportunities to the junior PhD students and post-doctoral fellows involved in the project. As previously described, the Innovation Board will also play a mentoring role for researchers in the project looking for professional development in the industrial sector.
- **Shared junior fellows**. The academic partners of the project have agreed to facilitate the long-term exchange of post-doctoral researchers who will spend one or two years in one lab and then will move on to another for the remaining time of their three-year contract.

Finally the project may **sponsor relevant workshops and training events**.

3.2.2 Dissemination

One central activity of BioPreDyn is the dissemination of scientific and technological knowledge that is generated by the project. The support by the EC will be acknowledged in any measure taken to disseminate the project results and to engage with the public and the media. A project website will be designed by the PM team, based on previous experience (see, for example, http://www.systemtb.eu, or http://www.geuvadis.eu), and the content will be implemented together with all partners. The BioPreDyn website will be the main portal for the scientific community, the general public, stakeholders and policy makers.

Dissemination of scientific and technological results to the scientific community and industry will be implemented following the usual procedures: publication in **peer-reviewed journals** (preferably open-access publications), presentations at **international conferences and professional trade shows**, practical courses, seminars and workshops. Moreover, the consortium will create a unified and consistent **code infrastructure** that includes all the methods and tools implemented and developed during the project. This platform will enable easy establishment of flexible, automated workflows, and guarantee interoperability and comparison of methods and tools. It will be distributed freely for academic purposes, and a commercial version will be developed by CSM (see section 3.2.3).

Overall, BioPreDyn will promote **synergies with other EU and non-EU funded initiatives** and European communication platforms, such as CommNet about food quality and safety (www.commnet.eu).

To catalyze the interaction between experiment and theory in the area of cellular network inference and quantitative model building in systems biology, the DREAM (Dialogues in Reverse Engineering Assessment of Methods; www.the-dream-project.org) initiative was launched six years ago. DREAM revolves around optimisation ‘challenges’ that are posed yearly to the community, which it then tries to solve; results are evaluated and discussed in a conference. The emerging picture is one where there is no single method that performs best for all types of data and questions; indeed, combining results of multiple methods, often leads to better results (Prill *et al.* 2010). Partners of our BioPreDyb are either already actively involved in DREAM (EMBL, FTELE.IGL) or will be encouraged to participate in it.

BioPreDyn partners will follow the activities and communications from the **European Technology Platform for Sustainable Chemistry** (SuSChem) and participate in their stakeholder workshops.

The CRG has recently submitted a proposal to the EC together with other European top research institutes (such as the Karolinksa Institute, EMBL, INSERM and Charité University) to create a European network for communication of scientific results funded by the EC, targeting a broad range of groups (general public, schools, policy makers, stakeholders, etc). If successful, the network will be highly beneficial for the communication strategy of BioPreDyn.

The project will also engage in a **dialogue with society at large and specific target groups**. The type of modelling and optimisation for reverse-engineering carried out in this project is applicable to a very wide range of complex problems (ranging from biological and biotechnological systems as those described above, to the modelling of ecosystems, to the modelling of complex organizations

and financial markets etc.). It is therefore, of broad importance to society. The project **webpage** will contain a session dedicated to the general public, which will explain the importance and context of our project, including videos and other source of media material. Project results will also be linked to the "Bulletin Board System" database of the Enterprise Europe Network, and communicated to the authorities managing the Cohesion Policy Fund.

Communication actions related to the project will be co-ordinated by the PM Team with the collaboration of the partners and the press offices at their respective institutions. Press releases concerning BioPreDyn will be co-ordinated and synchronized in the partner countries. Media articles and interviews in newspapers, radio, TV and podcasts will be promoted to enable dissemination to a broader audience and increase public engagement in scientific research and biotechnology.

3.2.3 Exploitation of Project Results and Management of Intellectual Property

As previously mentioned (see section 2.1.2), the project will produce intellectual property (IP) of significant value for the scientific community, for SMEs involved in bio-technological production processes, and for companies interested in using modelling for process optimisation in general. The effective management of IP is guaranteed by dedicated work packages (WP8 & WP10) and the creation of an Innovation Board (IB) formed by experts in technology transfer and software development from each partners' institution and from each of the SMEs. The IB will have a central role in the project from its beginning. Additionally, a **Consortium Agreement (CA)** will be signed by all academic and private partners, which will regulate all aspects of IP, access rights and software in detail.

The main exploitable IP expected to result from our project consists of software developed by all academic partners, and the SME CSM. Additionally, a second SME (INSIL) and several academic partners will be interested in incorporating methods developed during this project into their own (open-source or proprietary) software platforms. One of the main legal issues, therefore, concerns integration of different codes under different licenses into common computational frameworks. This issue will be dealt with by the Innovation Board (IB), as detailed in section 2.1.2 above.

As general philosophy, we will aim at implementing an open code-sharing environment within the consortium, in which academic partners agree to exchange code (wherever possible) or specifications of algorithms (in pseudo-code or equivalent formats) among themselves. Furthermore, code to be developed within this project will be offered to partners within the consortium before it is offered to outside companies interested in its commercial exploitation. Those partners who need or want to integrate codes into their respective computational frameworks will be granted a first option to negotiate an agreement to adapt the required code (generated by another partner) on conditions to be discussed on a case-by-case basis. Such agreements should result in royalty-bearing licenses if commercial exploitation is intended.

CSM operates a dual-licensing strategy, which it will implement for shared code developed within BioPreDyn. Their software will be made available to the academic community for free, while an enhanced commercial version will be available, featuring an improved GUI or other features concerning ease of use of the package.

This basic IP framework will be integrated and further elaborated in the Consortium Agreement, according to the DESCA model and including the special module with detailed provisions on software, which allocate liability and responsibility between the parties.

4. Ethics Issues

ETHICS ISSUES TABLE

(Note: Research involving activities marked with an asterisk * in the left column in the table below will be referred automatically to Ethics Review)

	Research on Human Embryo/ Foetus	YES	Page
*	Does the proposed research involve Human Embryos?	NO	
*	Does the proposed research involve Human Foetal Tissues/ Cells?	NO	
*	Does the proposed research involve Human Embryonic Stem Cells (hESCs)?	NO	
*	Does the proposed research on Human Embryonic Stem Cells involve cells in culture?	NO	
*	Does the proposed research on Human Embryonic Stem Cells involve the derivation of cells from Embryos?	NO	
	I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL	YES	

	Research on Humans	YES	Page
*	Does the proposed research involve children?	NO	
*	Does the proposed research involve patients?	NO	
*	Does the proposed research involve persons not able to give consent?	NO	
*	Does the proposed research involve adult healthy volunteers?	NO	
	Does the proposed research involve human genetic material?	NO	
	Does the proposed research involve human biological samples?	NO	
	Does the proposed research involve human data collection?	NO	
	I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL	YES	

	Privacy	YES	Page
	Does the proposed research involve processing of genetic information or personal data (e.g. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)?	NO	
	Does the proposed research involve tracking the location or observation of people?	NO	
	I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL	YES	

	Research on Animals [1]	YES	Page
	Does the proposed research involve research on animals?	NO	
	Are those animals transgenic small laboratory animals?	NO	
	Are those animals transgenic farm animals?	NO	
*	Are those animals non-human primates?	NO	
	Are those animals cloned farm animals?	NO	
	I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL	YES	

[1] The type of animals involved in the research that fall under the scope of the Commission's Ethical Scrutiny procedures are defined in the Council Directive 86/609/EEC of 24 November 1986 on the approximation of laws, regulations and administrative provisions of the Member States regarding the protection of animals used for experimental and other scientific purposes Official Journal L 358 , 18/12/1986 p. 0001 - 0028

	Research Involving ICP Countries[2]	YES	Page
	Is the proposed research (or parts of it) going to take place in one or more of the ICP Countries?	NO	
	Is any material used in the research (e.g. personal data, animal and/or human tissue samples, genetic material, live animals, etc):		
	a) Collected in any of the ICP Countries?	NO	
	b) Exported to any other country (including ICPC and EU Member States)?	NO	
	I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL	YES	

	Dual Use	YES	Page
	Research having direct military use	NO	
	Research having the potential for terrorist abuse	NO	
	I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL	YES	

[19] In accordance with Article 12(1) of the Rules for Participation in FP7, 'International Cooperation Partner Country (ICPC) means a third country which the Commission classifies as a low-income (L), lower-middle-income (LM) or upper-middle-income (UM) country. The list of countries is given in annex 1 of the work programme. Countries associated to the Seventh EC Framework Programme do not qualify as ICP Countries and therefore do not appear in this list.

5. Consideration of Gender Aspects

Equal Opportunity Policy

An equal opportunity policy regarding recruitment will be followed by all partners, without, however, taking precedence over quality and competence. Researchers will not be discriminated in any way on the basis of age, ethnic, national or social origin, religion or belief, sexual orientation, language, disability, political opinion, or economic condition ("*non-discrimination principle*").

The institution of the main project co-ordinator (CRG) has adhered to the "European Charter for Researchers and Code of Conduct for the Recruitment of Researchers" and will remind all partners of good practise for researchers and employers as stated in the EC Document.

Gender Balance

Europe is still far from gender balance in science and technology, especially in the leading and decision-making positions (*She Figures 2009*, European Commission). Similarly, the biotechnological sector witnesses the same underrepresentation of women in leadership roles (*EC-US Task Force on Biotechnological Research Workshop*, 2009). Within the consortium, one group leader is a woman (Joke Blom, CWI). The project will promote: i) gender awareness by collecting relevant documents, statistics, events, etc in a dedicated session of the website; ii) transparency in the selection procedures (in accordance to the section above); iii) mentoring of young researchers (female and male) in the development of their scientific career and especially in "making the jump" to independent positions.

Work & Life Balance

Most of the partner institutions promote gender-friendly policies with flexible working hours and appropriate infrastructures to help scientists reconcile professional and private life. BioPreDyn meetings and workshops will be organized during working days and we will do our best to provide childcare whenever needed. Moreover, the project will take a specific concrete action. Up to 2 awards of the value of 3,000 € each will be established for young researchers appointed by the project (independently of gender). These prizes shall be assigned in case of maternity/paternity and shall be used to top-up the salary of technicians, students, or post-doctoral research fellows to carry over the project of the mother/father scientist or as a contribution to baby-sitting/domestic help to help the mother/father scientist to go back to research. The candidates will be selected by the PSC based on scientific excellence. However, priority will be give to women.

6. Annexes

6.1 References

Allen M & Tildesley D (2002). *Computer Simulation of Liquids.* Oxford University Press.

Alper H, Jin YS, Moxley JF & Stephanopoulos G (2005). Identifying gene targets for the metabolic engineering of lycopene biosynthesis in *Escherichia coli*. *Metab Eng* 7:155–64.

Ashyraliyev M, Jaeger J & Blom JG (2008). Parameter estimation and determinability analysis applied to *Drosophila* gap gene circuits. *BMC Syst Biol* 2: 83.

Ashyraliyev M, Fomekong-Nanfack Y, Kaandorp JA & Blom JG (2009a). Systems biology: parameter estimation for biochemical models. *FEBS J* 276: 886–902.

Ashyraliyev M, Siggens K, Janssens H, Blom J, Akam M & Jaeger J (2009b). Gene circuit analysis of the terminal gap gene *huckebein*. *PLoS Comp Biol* 5: e1000548.

Asprey SP & Macchietto S (2000) Statistical tools for optimal dynamic model building. *Comput Chem Eng* 24: 1261–7.

Audoly S, Bellu G, D'Angio L, Saccomani MP & Cobelli C (2001). Global identifiability of nonlinear models of biological systems. *IEEE Trans Biomed Eng* 48:55–65.

Babuška I, Nobile F & Tempone R (2008). A systematic approach to model validation based on Bayesian updates and prediction related rejection criteria. *Comp Meth Appl Mech Eng.* 197: 2517–39.

Banga JR & Balsa-Canto E (2008). Parameter estimation and optimal experimental design. *Essays Biochem* 45: 195–210.

Bansal M, Belcastro V, Ambesi-Impiombato A & di Bernardo D (2007). How to infer gene networks from expression profiles. *Mol Syst Biol* 3: 78.

Balsa-Canto E, Alonso AA & Banga JR (2008). Computational procedures for optimal experimental design in biological systems. *IET Syst Biol* 2:163–72.

Balsa-Canto E, Alonso AA & Banga JR (2010). An iterative identification procedure for dynamic modeling of biochemical networks. *BMC Syst Biol* 4:11.

Bassingthwaighte JB, Raymond GM, Ploger JD, Schwartz LM & Bukowski TR (2006). GENTEX, a general multiscale model for *in vivo* tissue exchanges and intraorgan metabolism. *Phil Trans Roy Soc A* 364: 1423–42.

Bauer I, Bock HG, Korkel S & Schloder JP (2000). Numerical methods for optimum experimental design in DAE systems. *J Comp Appl Math* 120, 1–25.

Becskei A, Kaufmann B & van Oudenaarden A (2005). Contributions of low molecule number and chromosomal positioning to stochastic gene expression (2005). *Nat Genet* 37:937–44.

Bellu G, Saccomani MP, Audoly S & D'Angiò L (2007). DAISY: a new software tool to test global identifiability of biological and physiological systems. *Comp Meth Prog Biomed* 88: 52–61.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FCP, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J & Vingron M (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat Genet* 29 365–71.

Breiman L & Friedman J (1985). Estimating optimal transformations for multiple regression and correlation. *J Am Stat Assoc* 80: 580–98.

Bro C, Regenberg B, Förster J & Nielsen J (2006). In silico aided metabolic engineering of *Saccharomyces cerevisiae* for improved bioethanol production. *Metab Eng* 8:102–11.

Broughton JQ (1999). Concurrent coupling of length scales: methodology and application, *Phys Rev B* 60: 2391–403.

Casey FP, Baird D, Feng Q, Gutenkunst RN, Waterfall JJ, Myers CR, Brown KS, Cerione, RA & Sethna JP (2007). Optimal experimental design in an epidermal growth factor receptor signalling and down-regulation model. *IET Syst Biol* 1: 190–202.

Cedersund G & Roll J (2009). Systems biology: model based evaluation and comparison of potential explanations for given biological data. *FEBS J* 276: 903–22.

Chen WW, Schoeberl B, Jasper PJ, Niepel M, Nielsen UB, Lauffenburger DA & Sorger PK (2009) Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. *Mol Syst Biol* 5: 239.

Dada JO, Spasic I, Paton NW & Mendes P (2010). SBRML: a markup language for associating systems biology data with models. *Bioinformatics* 26: 932–38.

Deb K (2009). *Multi-objective optimization using evolutionary algorithms*. Wiley & Sons.

de Smet R & Marchal K (2010). Advantages and limitations of current network inference methods. *Nat Rev Microbiol* 8: 717–29.

Dobrzyński M, Vidal Rodriguez J, Kaandorp J & Blom J (2007). Computational methods for diffusion-influenced biochemical reactions. *Bioinformatics* 23:1969–77.

Dobrzyński M (2011). *Molecules in Motion: a theoretical study of noise in gene expression and cell signalling.* PhD thesis, University of Amsterdam.

Dobson PD, Smallbone K, Jameson D, Simeonidis E, Lanthaler K, Pir P, Lu C, Swainston N, Dunn WB, Fisher P, Hull D, Brown M, Oshota O, Stanford NJ, Kell DB, King RD, Oliver SG, Stevens RD & Mendes P.(2010). Further developments towards a genome-scale metabolic model of yeast. *BMC Syst Biol* 4:145.

Esmaeili A & Jacob C (2009). A multi-objective differential evolutionary approach toward more stable gene regulatory networks. *BioSystems* 98: 127–36

Feng XJ & Rabitz H: Optimal identification of biochemical reaction networks. *Biophys J* 2004, 86:1270–81.

Feng X-J, Rabitz H, Turinici G, & Le Brisn C (2006). A closed-loop identification protocol for nonlinear dynamical systems. *J Phys Chem* A: 110: 7755–62.

Finnerty JR, Pang K, Burton P,Paulson D & Martindale MQ (2004). Origins of bilateral symmetry: Hox and Dpp expression in a sea anemone. *Science* 304: 1335–7.

Fomekong Nanfack Y, Kaandorp JA & Blom JG (2007). Efficient parameter estimation for spatio-temporal models of pattern formation: Case study of *Drosophila melanogaster. Bioinformatics* 23, 3356–63.

Friedman N, Linial M, Nachman I & Pe'er D (2000). Using Bayesian networks to analyze expression data. *J Comp Biol* 7: 601–20.

Gadkar KG, Gunawan R & Doyle FJ (2005). Iterative approach to model identification of biological networks. *BMC Bioinformatics*, 6: 155.

Gardiner CW (1983). *Handbook of Stochastic Methods.* Berlin*:* Springer-Verlag.

Gelman A., Carlin J, Stern H & Rubin D (1995). *Bayesian Data Analysis*. Chapman & Hall.

Geisser S (1993). *Predictive inference: an introduction.* Chapman & Hall.

Godfrey KR & *Fitch* WR (1984). The deterministic identifiability of nonlinear pharmaco-kinetic models. *J Pharmacokinet Biopharm* 12: 177–91.

Goldenfeld and Kadanoff (1999) Simple lessons from Complexity. *Science* 284: 87–9.

Guo H, Meng Y & Jin Y (2009). A cellular mechanism for multi-robot construction via evolutionary multi-objective optimization of a gene regulatory network. *BioSystems* 98: 193–203.

Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR & Sethna JP (2007). Universally sloppy parameter sensitivities in systems biology models. *PLoS Comp Biol* 3: e189.

Halford SE & Marko JF (2004). How do site-specific DNA-binding proteins find their targets? *Nucl Acid Res* 32*:* 3040–52.

Handl J, Kell DB & Knowles J (2007). Multiobjective Optimization in Bioinformatics and Computational Biology. *IEEE/ACM Trans Comp Biol Bioinf* 4: 279–92.

Hengl S, Kreutz C, Timmer J & Maiwald T (2007). Data-based identifiability analysis of non-linear dynamical models. *Bioinformatics* 23, 2612–18.

Herrgard MJ, Swainston N, Dobson P, Dunn WB, Arga KY, Arvas M, Blüthgen N, Borger S, Costenoble R, Heinemann M, Hucka M, Le Novère N, Li P, Liebermeister W, Mo ML, Oliveira AP, Petranovic D, Pettifer S, Simeonidis E, Smallbone K, Spasić I, Weichart D, Brent R, Broomhead DS, Westerhoff HV, Kirdar B, Penttilä M, Klipp E, Palsson BØ, Sauer U, Oliver SG, Mendes P, Nielsen J & Kell DB (2008). A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature Biotechnol* 26: 1155–60.

Hidalgo ME & Ayesa E (2001). Numerical and graphical description of the information matrix in calibration experiments for state-space models. *Wat Res* 35: 3206–14.

Hlavacek WS, Faeder JR, Blinov ML, Posner RG, Hucka M & Fontana W (2006). Rules for modeling signal-transduction systems. *Sci STKE:* re6.

Hoekstra AG (2010). Multiscale coupling of a lattice Boltzmann simulation of blood flow to cell- and tissue-level processes: the case of in-stent restenosis; in Pereira JCF & Sequeira A (eds), *Proc V Europ Conf Comp Fluid Dyn*, Lisbon, Portugal.

Honkela A, Girardot C, Gustafson EH, Liu Y-H, Furlong EEM, Lawrence ND & Rattray M (2010) Model-based method for transcription factor target identification with limited data. *Proc Natl Acad Sci USA* 107: 7793–8.

Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H & the SMBL Forum (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19: 524–31

Jaeger J, Surkova S, Blagov M, Janssens H, Kosman D, Kozlov KN, Myansikova E, Vanario-Alonso CE, Samsonova M, Sharp DH & Reinitz J (2004a). Dynamic control of positional information in the early *Drosophila* embryo. *Nature* 430: 368–71.

Jaeger J, Blagov M, Kosman D, Kozlov KN, Manu, Myasnikova E, Surkova S, Vanario-Alonso, CE, Samsonova M, Sharp DH & Reinitz J (2004b). Dynamical analysis of regulatory interactions in the gap gene system of *Drosophila melanogaster. Genetics* 167: 1721–37.

Jaqaman K & Danuser G (2006). Linking data to models: data regression. *Nat Rev Mol Cell Bio* 7: 813–9.

Karniakidis GE & Glimm J (eds) (2006). *Uncertainty quantification in simulation science*. *J Comp Phys* 217 (Special Issue).

Kitano H (2002). Systems biology: a brief overview. *Science* 295:1662–4.

Kremling A, Fischer S, Gadkar K, Doyle FJ, Sauter T, Bullinger E, Allgower F & Gilles ED (2004). A benchmark for methods in reverse engineering and model discrimination: problem formulation and solutions. *Genome Res.* 14: 1773–85.

Kreutz C & Timmer J (2009). Systems biology: experimental design. *FEBS J* 276: 923–42.

Kusserow A, Pang K & Sturm C, Hrouda M, Lentfer J, Schmidt HA, Technau U, von Haeseler A, Hobmayer B, Martindale MQ & Holstein TW (2005). Unexpected complexity of the Wnt gene family in a sea anemone. *Nature*, 433: 156–60.

Kutalik Z, Cho K-H & Wolkenhauer O (2004). Optimal sampling time selection for parameter estimation in dynamic pathway modeling. *BioSystems*, 75: 43–55.

Lawrence ND (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *J Mach Learn Res* 6: 1783–816.

Lawrence N, Girolami M, Rattray M & Sanguinetti G (eds.) (2010). *Learning and Inference in Computational Systems Biology.* MIT Press, Cambridge, MA.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne J-B, Volkert TL, Fraenkel E, Gifford DK & Young RA (2002). Transcriptional regulatory networks in *Saccharomyces cervisiae*. Science 298: 799–804.

Li X-Y, MacArthur S, Bourgon R, Nix D, Pollard DA, Iyer VN, Hechmer A, Simirenko L, Stapleton M, Luengo Hendriks CL, Chu HC, Ogawa N, Inwood W, Sementchenko V, Beaton A, Weiszmann R, Celniker SE, Knowles DW, Gingeras T, Speed TP, Eisen MB & Biggin MD (2008). Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biol* 6: e27.

Liebermeister W, Uhlendorf J & Klipp E (2010). Modular rate laws for enzymatic reactions: thermodynamics, elasticities and implementation *Bioinformatics* 26: 1528–34.

Ljung L (1999). *System Identification: Theory for the User*. Prentice Hall.

Manu, Surkova S, Spriov AV, Gursky VV, Janssens H, Kim A-R, Radulescu O, Vanario-Alonso CE, Sharp DH, Samsonova M & Reinitz J (2009). Canalization of gene expression in the *Drosophila* blastoderm by gap gene cross regulation. *PLoS Biol* 7: e1000049.

Manu, Surkova S, Spriov AV, Gursky VV, Janssens H, Kim A-R, Radulescu O, Vanario-Alonso CE, Sharp DH, Samsonova M & Reinitz J (2009). Canalization of gene expression and domain shifts in the *Drosophila* blastoderm by dynamical attractors. *PLoS Comp Biol* 5: e1000303.

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R & Califano A (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7 (Suppl. 1): S7.

Melas (2006). *Functional approach to optimal experimental design.* Springer Verlag.

Mendes P & Kell DB (1998). Non-linear optimisation of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics* 10, 869–83.

Metzler R (2001). The future is noisy: the role of spatial fluctuations in genetic switching. *Phys Rev Lett* 87: 068103.

Mjolsness E, Sharp DH & Reinitz J (1991). A connectionist model of development. *J Theor Biol* 152: 429–53.

Moles C, Mendes P & Banga JR (2003). Parameter estimation in biochemical pathways: a comparison of global optimisation methods. *Genome Res* 13: 2467–74.

Morris MK, Saez-Rodriguez J, Sorger PK & Lauffenburger DA (2010). Logic-based models for the analysis of cell signaling networks. *Biochemistry* 49 3216–24.

Oberkampf WL & Barone MF. Measures of agreement between computation and experiment: validation metrics. *J Comp Phys* 217: 5–36.

Oden T, Moser M & Ghattas O (2010). Computer predictions with quantified uncertainty, part I. *SIAM News* 43(9).

Oden T, Moser M & Ghattas O (2010). Computer predictions with quantified uncertainty, part II. *SIAM News* 43(10).

Okino M & Mavrovouniotis M (1998) Simplification of mathematical models of chemical reaction systems. *Chem Rev* 98: 391–408.

Palsson BO & Thiele I (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protoc.* 5: 93–121.

Pearson R, Liu X Sanguinetti G, Milo M, Lawrence ND & Rattray M (2009). Puma: a Bioconductor package for propagating uncertainty in microarray analysis" *BMC Bioinformatics* 10: 211.

Perkins TJ, Jaeger J, Reinitz J & Glass L (2006). Reverse Engineering the Gap Gene Network of *Drosophila melanogaster. PLoS Comp Biol* 2: e51.

Pohjanpalo J (1978) System identifiability based on the power series expansion of the solution. *Math Biosci* 41: 21–33.

Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopolous LG, Xue X, Clarke ND, Altan-Bonnet G & Stolovitzky G (2010). Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS ONE* 5: e9202.

Radulescu O, Gorban AN, Zinovyev A & Lilienbaum A (2008) Robust simplifications of multiscale biochemical networks. *BMC Syst Biol* 2: 86.

Reinitz J & Sharp DH (1995). Mechanism of *eve* stripe formation. *Mech Dev* 49: 133–58.

Robert, CP (2007). *The Bayesian Choice*. Springer Verlag.

Rodriguez-Fernandez M, Mendes P & Banga JR (2006a). A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *BioSystems* 83: 248–65.

Rodriguez-Fernandez M, Egea JA & Banga JR (2006b). Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. *BMC Bioinformatics* 7: 483.

Saez-Rodriguez J, Goldsipe A, Muhlich J, Alexopoulos LG, Millard B, Lauffenburger DA, Sorger PK (2008). Flexible informatics for linking experimental data to mathematical models via DataRail. *Bioinformatics* 24:840–7.

Saez-Rodriguez J, Alexopoulos LG, Epperlein J, Samaga R, Lauffenburger DA, Klamt S & Sorger PK (2009). Discrete logic modeling as a means to link protein signaling networks with functional analysis of mammalian signal transduction. *Mol Syst Biol* 5: 331.

Sahle S, Mendes P, Hoops S & Kummer U (2008). A new strategy for assessing sensitivities in biochemical models. *Phil Trans Roy Soc A* 366: 3619–31.

Schittkowski K (2002). *Numerical Data Fitting in Dynamical Systems*. Kluwer.

Sloot PMA & Hoekstra AG (2010). Multi-scale modelling in computational biomedicine. *Brief Bioinf* 11: 142–52.

Smallbone K, Simeonidis E, Broomhead DS & Kell DB (2007). Something from nothing: bridging the gap between constraint-based and kinetic modelling. *FEBS J* 274: 5576–85.

Smallbone K, Simeonidis E, Swainston N & Mendes P (2010). Towards a genome-scale kinetic model of cellular metabolism. *BMC Syst Biol* 4: 6.

Stelling J (2004). Mathematical models in microbial systems biology. *Curr Op Microbiol* 7: 513–518.

Tamulonis C, Postma M, Marlow H, Magie C, de Jong J & Kaandorp JA (2011). Morphometrics and modeling of *Nematostella vectensis* gastrulation. *Dev Biol* (in press).

Tsai K & Wang F (2005) Evolutionary optimisation with data collocation for reverse engineering of biological networks. *Bioinformatics* 21: 1180–8.

Vajda S, Godfrey KR & Rabitz H (1989). Similarity transformation approach to structural identifiability of nonlinear models. *Math Biosci* 93: 217–48.

van Kampen NG (1997)*. Stochastic processes in physics and chemistry,* Elsevier*.*

van Someren EP, Wessels LFA, Backer E & Reinders MJT (2003). Multi-criterion optimization for genetic network modeling. *Signal Proc* 83: 763–75.

Vyshemirsky V & Girolami MA (2008). Bayesian ranking of biochemical system models. *Bioinformatics* 24: 833–9.

Walter, E. & Pronzato, L. (1997) Identification of Parametric Models from Experimental Data. Springer.

Wolkenhauer O & Mesarović M (2005). Feedback dynamics and cell function: Why systems biology is called systems biology. *Mol BioSyst* 1: 14–6.

Zwolak J, Tyson J, Watson L (2005). Globally optimised parameters for a model of mitotic control in frog egg extracts. *IEE Proc Syst Biol* 152: 81–92.