Theoretical Biology: Modeling and Simulation of Biological Systems and Laboratory Methods

INAUGURAL-DISSERTATION

zur Erlangung des akademischen Grades des
Doktors der Naturwissenschaften (Dr. rer. nat.)

eingereicht im Fachbereich Biologie, Chemie, Pharmazie
der Freien Universität Berlin

vorgelegt von

CHRISTOPH WIERLING

aus Münster, Deutschland

November, 2008

Die vorliegende Arbeit wurde in der Zeit von August 1999 bis November 2008 am Max-Planck-Institut für molekulare Genetik in Berlin-Dahlem in der Abteilung von Herrn Prof. Dr. Hans Lehrach in der Arbeitsgruppe von Herrn Dr. Ralf Herwig angefertigt.

1. Gutachter:	Prof. Dr. Hans Lehrach Max-Planck-Institut für Molekulare Genetik
2. Gutachter:	Prof. Dr. Volker Erdmann Freie Universität Berlin

Disputation am 26. Mai 2009

Contents

List of Figures

List of Tables

Summary

Mathematical modeling and simulation techniques have turned out to be valuable tools for the understanding of complex systems in different areas of research and engineering. In recent years this approach came to application frequently also in biology resulting in the establishment of the research area systems biology. Systems biology tries to understand the behavior of complex biological systems by means of mathematical approaches. This requires the integration of qualitative and quantitative experimental data into coherent models. Currently, systems biology usually investigates biochemical reaction networks of cellular systems. A challenging task is the construction of large models that requires computer-assisted data integration, simulation and evaluation.

In this work I have elaborated technical bases for the computer-assisted modeling of biological systems and experimental techniques. For this I have developed the program PyBioS that provides a user-friendly Web application (`http://pybios.molgen.mpg.de`) and brings in automation for several important tasks required for the development, implementation, and simulation of cellular models. For the description of cellular reaction systems PyBioS makes use of object-oriented programming, well established methods for the mathematical description of biochemical reaction systems based on ordinary differential equation systems, and novel interfaces to biochemical pathway databases (e.g., Reactome, KEGG). In addition PyBioS provides several different functions for the analysis and visualization.

The benefit obtained by mathematical modeling of biological systems using PyBioS is illustrated for segmentation of the body (somitogenesis) as, e.g., taking place during embryogenesis. The parameterized somitogenesis model I have developed comprises three signaling pathways, namely Notch, Wnt, and FGF that are known to be relevant for somitogenesis. The model shows a regular oscillation controlled by extracellular Wnt3a. Below a critical threshold concentration of Wnt3a the oscillation that is controlled by Wnt signaling arrests and approaches a steady state. These findings are conform to experimental observations found during determination of somite boundaries.

Besides the analysis of biological systems, modeling strategies can also be used for the evaluation of biotechnological experimental techniques. To study this I have perfomed simulations of DNA array hybridization experiments for the evaluation of critical parameters during subsequent image and data analysis. Therefore I have carried out simulation studies on several error parameters arising in complex hybridization experiments, such as spot shape, spot position and background noise. My results show how measurement errors can be balanced by the analysis tools.

Zusammenfassung (German Summary)

In verschiedenen Bereichen der Natur- und Ingenieurswissenschaften hat sich die mathematische Modellierung als ein geeignetes Werkzeug erwiesen, um komplexe Systeme besser zu verstehen. Dieser Ansatz findet auch immer häufiger Anwendung in der Biologie und führte zur Etablierung der Systembiologie. Die Systembiologie versucht mit Hilfe mathematischer Ansätze das komplexe Verhalten biologischer Systeme besser zu verstehen. Dies erfordert die Integration qualitativer und quantitativer Daten in kohärente Modelle. Derzeit werden in der Systembiologie häufig biochemische Reaktionsnetzwerke zellulärer Systeme betrachtet. Eine besondere Herausforderung stellt dabei die Modellierung grosser Systeme dar, die eine massive, computergestützte Datenintegration, Simulation und Auswertung erfordert.

In dieser Arbeit habe ich Grundlagen für die computergestützte Modellierung biologischer Systeme und experimenteller Verfahren erarbeitet. Das von mir hierfür entwickelte Programm PyBioS bietet eine benutzerfreundliche Web-Schnittstelle (`http://pybios.molgen.mpg.de`) und automatisiert viele Schritte, die für die Erstellung, Implementierung und Simulation zellulärer Modelle erforderlich sind. Für die Beschreibung der Modelle wurden dabei objektorientierte Ansätze der Informatik, etablierte Methoden der Modellierung biochemischemischer Reaktionssysteme basierend auf gewöhnlichen Differentialgleichungssystemen, sowie neuartige Schnittstellen zu Datenbanken biochemischer Reaktionswege (z.B. Reactome, KEGG) genutzt bzw. implementiert. Zudem bietet PyBioS verschiedene Funktionalitäten für die Analyse und Visualisierung.

Unter Verwendung von PyBioS wird am Beispiel der embryonalen Segmentierung (Somitogenese) gezeigt, wie mathematische Modellierung zum Verständnis biologischer Systeme beitragen kann. Das von mir entwickelte parametrisierte Modell umfasst die Signalwege Notch, Wnt und FGF, von denen bekannt ist, dass sie an der Determinierung der Somitenbildung beteiligt sind. Das Modell zeigt eine von extrazellulärem Wnt3a kontrollierte Oszillation. Unterhalb einer kritischen Wnt3a Konzentration bricht die vom Wnt Signalweg kontrollierte Oszillation ab und geht in einen stationären Zustand über, der den Beobachtungen für die Determination einer Somitengrenze entspricht.

Neben der Analyse biologischer Systeme kann Modellierung auch für die Evaluation biotechnologischer, experimenteller Methoden genutzt werden. Dies wurde für DNA-Array Hybridisierungsexperimente genauer untersucht. Anhand simulierter Daten wurden kritische Parameter der anschliessenden Bild- und Datenanlayse bewertet. Hierfür habe ich Simulationsstudien verschiedener experimenteller Parameter komplexer Hybridisierungsexperimente, wie z.B. der Spot-Form und Spot-Position, oder dem Hintergrundrauschen, durchgeführt. Meine Ergebnisse zeigen, wie Messfehler anhand geeingeter Analyseprogramme kompensiert werden können.

1 Introduction

For a long time research in molecular biology has been focussed on the analysis of specific components of the cellular network (genes, proteins, metabolites) one by one. By this approach thousands of genes have successfully been characterised and functionally annotated. But biological systems are complex and their characteristics are a result of a highly interwoven interaction network continuously developing through time and space. Fundamental characteristics of living systems, like the assimilation of nutrients, growth and reproduction, or the perception of (environmental) signals and their processing can be narrowed down basically to a single unit all living things are composed of: the cell (Schwann and Schleiden, 1839, 1847). Thus, the understanding of the characteristics of cellular systems is essential, but it requires an approach that takes into account both interactions on the molecular level as well as physiological functions that are characteristics of the whole organism. In particular in the light of understanding developmental processes or multigenic and complex diseases that cannot be pinned down to a single gene or component systems approaches become increasingly important.

During the last decade this gave rise to a new research area in biology called systems biology. Systems biology explanations of physiology and disease should be multi-level (Noble, 2002b); from molecular pathways and regulatory networks, through cells and organs, ultimately to the level of the whole organism or even to an ecosystem. With the use of computer models for such processes *in silico* predictions can be generated on the state of the disease or the effect of the individual therapy (Kitano, 2002; Herwig and Lehrach, 2006). Models are partial representations and their aim is to explain which features of a system are necessary and sufficient to understand it (Noble, 2002b). The performance of a model is mainly defined by its predictive power.

Systems biology is going to revolutionise our knowledge of disease mechanisms and the interpretation of data from high-throughput technologies. Systems biology is the coordinated study of biological systems by (1) investigating the components of cellular networks and their interactions, (2) applying experimental high-throughput and whole-genome techniques, and (3) integrating computational methods with experimental efforts (Klipp et al., 2005). This approach requires an integration of experimental and computational methods and, thus, an iterative process of data mining and data gathering (e.g., from scientific literature, databases and experiments), data integration, computational modeling and analysis, and finally validation of specific observations that were not explainable beforehand (Kitano, 2002).

Using data mining steps, one agglomerates sufficient details for the generation of model prototypes of the biological system under investigation. Eventually, using analysis methods, the mathematical model is refined, cross-validated with regard to internal and external features, for example using parameter estimation (Moles et al., 2003), and it is used to formulate new hypotheses that in turn are subject to further experimental investigation.

Systems biology methodology and approaches evolved rapidly in the last years driven by the new high-throughput technologies. A significant impulse was given by the large sequencing projects, such as the human genome project, which resulted in the nearly complete sequence of the human and other genomes (Lander et al., 2001; Venter et al., 2001). This knowledge builds the theoretical basis to compute gene regulatory motifs, to determine the exon-intron structure of genes and to derive the coding sequence of potentially all genes of many organisms. From the genome sequences probes for whole genome DNA arrays have been constructed that allow to monitor the transcriptome level of most genes active in a given cell- or tissue type. Proteomics technologies have been used to identify translation status on a large scale (2D-gels [Klose, 1975; Klose et al., 2002], mass spectrometry, reverse phase protein arrays [Paweletz et al., 2001]). Protein-protein interaction data involving thousands of components were measured to determine information on the proteome level (von Mering et al., 2002). Multiple databases of diverse aspects of biological systems exist[1], a variety of experimental techniques have produced gene and proteome expression data from various tissues and samples and important disease-relevant pathways have been investigated. Information on promoter regions and transcription factors is available for nearly all genes. This information - although extremely helpful - cannot be utilised sufficiently, because of the lack of integrative analysis tools. To validate such data in the system-wide hierarchical context ranging from DNA to RNA to protein to interaction networks and further on to cells, tissues, organs or even the whole individual, one needs to correlate and integrate such information. Thus, an important part of systems biology is data integration that provides a foundation for the development of computational models.

As mentioned above models should be generated on a multi-level basis, but need to be grounded on the molecular- and cellular-level so that a continuous spectrum of knowledge can be established. The question of the most suitable approach to system-level understanding has been addressed by Noble (2002b, 2006). He discussed the ’bottom-up’ and ’top-down’ approach to understand biological systems. The bottom-up approach starts with all the individual genes, proteins, metabolites, etc. and their individual reactions and interactions to

[1] Pathguide: The Pathway Resource List: `http://www.pathguide.org`

come up with an integrated molecular model for the prediction of general system properties. On the other hand, the top-down approach starts with the overall behaviour of systems (as in classical physiology with the analysis of the circulatory system, the respiratory, the immunological, and so on) and then progressively identifies and explores the elements of each system so as to deduce the underlying functions (Noble, 2006, p. 75). Both approaches have their strengths and limitations that lead to the ’middle-out’ approach originally proposed by Sydney Brenner and adopted by Denis Noble (2006, p. 79). It states the simple and pragmatic concept of starting at any level as long as enough data is available to feed into a simulation for the purpose of systems analysis. However, a crucial point is that models are always partial representations and their aim is explanation: to show which features of a system are necessary and sufficient to understand it (Noble, 2002b).

Modeling and simulation techniques are valuable tools for the understanding of complex biological systems. A computational approach offers the possibility to use simulations for the prediction of the dynamical behavior of biological systems according to the defined models, and to test the validity of the underlying assumptions (Kitano, 2002). To this end, it is necessary to construct computer-executable models that are consistent with experimental observations. The development of such a model is an iterative process of (1) model design based on existing knowledge, (2) simulation and model-analysis, which results in (3) the generation of new hypotheses that can be proven by experiments in the wet lab and used anew for model-refinement. This hypothesis-driven approach based on *in silico* experiments will support the experimental design or help to investigate questions that are not accessible to experimental inquiry. Noble (2002a) states that "physiological analysis requires an understanding of functional interactions between the key components of cells, organs and systems, as well as how these interactions change in disease states". He argues that there is no alternative to copying nature and computing these functional interactions to determine the logic of healthy and disease states.

1.1 Outline

In this work I present different applications of modeling in biology and biological research (Fig. 1.1). In the following sections I will outline the modeling of biological systems and discuss modeling tools currently used in systems biology (see Section 1.3). Later, in the Results, I will introduce the modeling and simulation system PyBioS, which I have developed in the course of this thesis (see Section 2.1) and deployed to different biological problems.

In particular, the PyBioS modeling system was used to build a model on somitogenesis (described in Section 2.2), which is a fundamental process during vertebrate development.

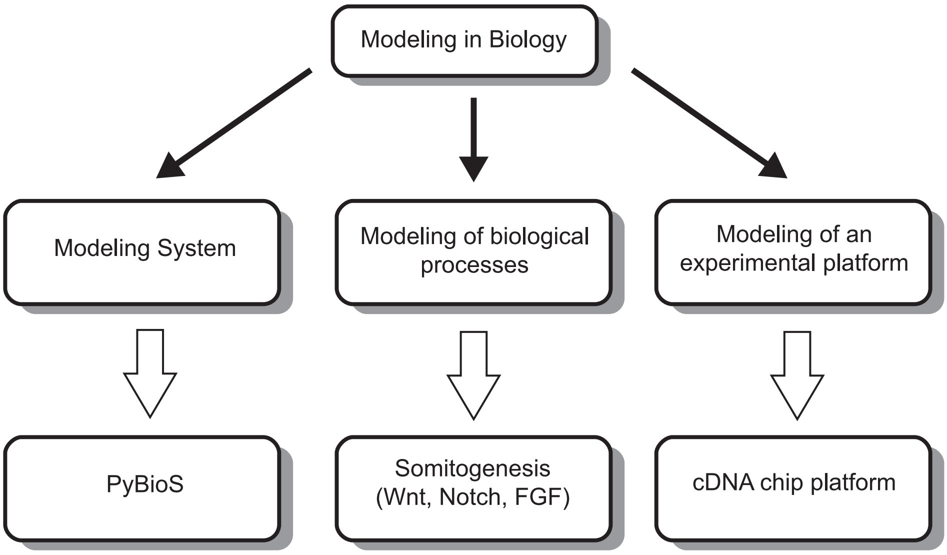

Modeling in Biology
Modeling System
Modeling of biological processes
Modeling of an experimental platform
PyBioS
Somitogenesis (Wnt, Notch, FGF)
cDNA chip platform

Figure 1.1: Overview of the thesis.

The model captures central components known or assumed to be involved in somitogenesis (that is introduced in Section 1.2.1). The model takes into account three signaling pathways triggered by signaling of Notch (Section 1.2.2.1), Wnt (Section 1.2.2.2) and Fgf (Section 1.2.2.3), as well as subsequent genes known to be regulated by these pathways.

Furthermore, I have applied modeling strategies to the evaluation of an experimental technique used in modern molecular biology. As, for example, common in engineering, modeling of technical processes can also help significantly by the evaluation of experimental platforms. In Section 1.4.1 I will introduce DNA arrays that became a common standard for expression profiling in molecular biology and in Section 2.3 I will evaluate error sources subject to cDNA arrays by the use of a computational model.

1.2 Biological Systems

Coordinated interactions between the different cellular components give rise to the astonishing complex but well coordinated processes of living organisms, such as the development of a multicellular organism (cf. Gilbert, 2003). Fundamental for development is the differentiation—the structural and functional specialisation of cells and tissues during ontogenesis. A first step during the differentiation process of higher animals is the formation of the germ layers ectoderm, endoderm, and mesoderm during gastrulation. Later on during early embryogenesis many animal species undergo a segmentation of the body axis. In vertebrates this segmentation is called somitogenesis and the segments that are formed during this process are the somites. In the following I will give a brief introduction on somitogenesis and a more detailed description of the molecular pathways that control this developmental process.

1.2.1 Somitogenesis

During gastrulation the embryo shapes into three germ layers: (1) the endoderm, the precursor of the gut and its associated glands, (2) the mesoderm, the precursor of the skeleton, smooth muscle, connective tissue, and vascular system, and (3) the ectoderm, the precursor of the epidermis and the nervous system. Following gastrulation, the dorsoventral axis is specified by signals from the node (which is the homolog in mouse and chicken of the frog Organizer, a region of the dorsal lip of the blastopore that is known for its crucial role in organizing the formation of the main body axis). During this process the ectoderm thickens, rolls up, and pinches off to form the neural tube and neural crest. Below the neural tube a rod of specialized cells derived from the mesoderm called the notochord elongates and forms the central axis of the embryo. On both sides of the notochord the unsegmented paraxial mesoderm or presomitic mesoderm (PSM) is formed that gets segmented later on in an anterior-to-posterior sequence while the embryo elongates at the tail bud (Alberts et al., 2008). During this segmentation process that is called somitogenesis small epithelial spheres, the somites, form along the length of the embryo (Fig. 1.2A). The somites eventually give rise to the vertebrae and ribs, the dermis of the dorsal skin, the muscles of the back, and the skeletal muscles of the body wall and limbs (Gilbert, 2003, p. 466). The number of somites and time period of their formation is highly constrained within a given species, but varies widely between different species (cf. Tab. 1.1). The final number of somites ranges from less than 50 (in a frog or a bird) to more than 300 (in a snake) (Alberts et al., 2008).

Table 1.1: Specific values on somitogenesis for different organisms. [1]Stickney et al. (2000); [2]Gilbert (2003); [3]Tam (1981)

Organism	Number of somites	Duration for a single somite formation
zebrafish	about 30[1]	ca. 30 minutes
chicken	50[2]	ca. 90 minutes
mouse	65[2,3]	ca. 120 minutes

Major components of somitogenesis are periodicity, epithelialization, specification, and differentiation. In mouse embryos the first somites form in the posterior headfold region around embryonic day 7.75 (E7.75). Subsequently, new somites arise at regular intervals in a strict anterior-to-posterior sequence from the unsegmented PSM (Hofmann et al., 2004).

The molecular process underlying somitogenesis has been studied in detail. It has been shown that targeted inactivation of *Notch* and *Delta* in mice leads to an impairment in somitogenesis (Conlon et al., 1995; de Angelis et al., 1997). This suggests that Notch signaling is involved in somitogenesis.

The periodic formation of equally sized somites implicates that a molecular, gene-regulatory oscillator is involved in somitogenesis. The first gene identified to oscillate during somitogenesis in chicken embryos was *c-hairy1* (Palmeirim et al., 1997). A second avian *hairy*-related gene found to cycle in the PSM is *c-hairy2*, which is closely related to mammalian gene *Hes1* (Jouve et al., 2000). *Hes1* is described as a downstream target of Notch signaling (Kageyama and Ohtsuka, 1999). It has a basic helix-loop-helix (bHLH) motif and acts as a transcriptional repressor (Sasai et al., 1992). Subsequently, several other genes showing a cyclic behaviour during somitogenesis were identified in fish, frog and mouse, implicating that the oscillator is conserved in vertebrates (Dequéant and Pourquié, 2008).

In 2001 Bessho et al. cloned another downstream Notch effector termed *Hes7* from mouse. *Hes7* has also a bHLH motif and was revealed to be specifically expressed in the PSM in a dynamic manner. *Hes7* was found to be controlled by Notch signaling and to encode also a transcriptional repressor (Bessho et al., 2001).

Another gene that has been identified to be required for the processing of *Notch1* and *Dll1* (*Delta* ligand) in the paraxial mesoderm is *Presenilin1* (Wong et al., 1997). Moreover, in the chicken embryo *lunatic fringe* (*Lnfg*), which encodes a glycosyltransferase that can modify the Notch receptor, has been shown to be activated periodically by Notch signaling in the PSM (Dale et al., 2003). Overexpressing *Lnfg* in the paraxial mesoderm abolishes the expression of cyclic genes including endogenous *Lnfg* itself and leads to defects in segmentation (Dale et al., 2003).

In zebrafish, all the cyclic genes identified so far belong to the Notch pathway. In amniotes (reptiles, birds and mammals), also genes of Wnt signaling and FGF signaling have been identified to oscillate in the PSM with periods correlating with the time for somite formation. In mouse, *Axin2*, a key negative feedback inhibitor of the Wnt pathway (see Section 1.2.2.2), has been found to show an analogous cyclic behavior. Moreover, observations for the hypomorphic *Wnt3a* mutation vestigial tail (*vt*) in mice implicate an involvement of Wnt signaling in somitogenesis (Aulehla et al., 2003).

In a large scale microarray study conducted by Dequéant et al. (2006) multiple genes have been identified that show an oscillatory behavior in the PSM during somitogenesis. When ordered by their time of maximum expression in the segmentation clock cycle, the cyclic genes could be assigned to two mutually exclusive main clusters with opposite phase. One of the clusters contains known cyclic genes regulated by Notch and FGF signaling and the other includes those controlled by Wnt signaling (Tab. 1.2).

A first theoretical model for the sequential positioning of somites was introduced by Cooke

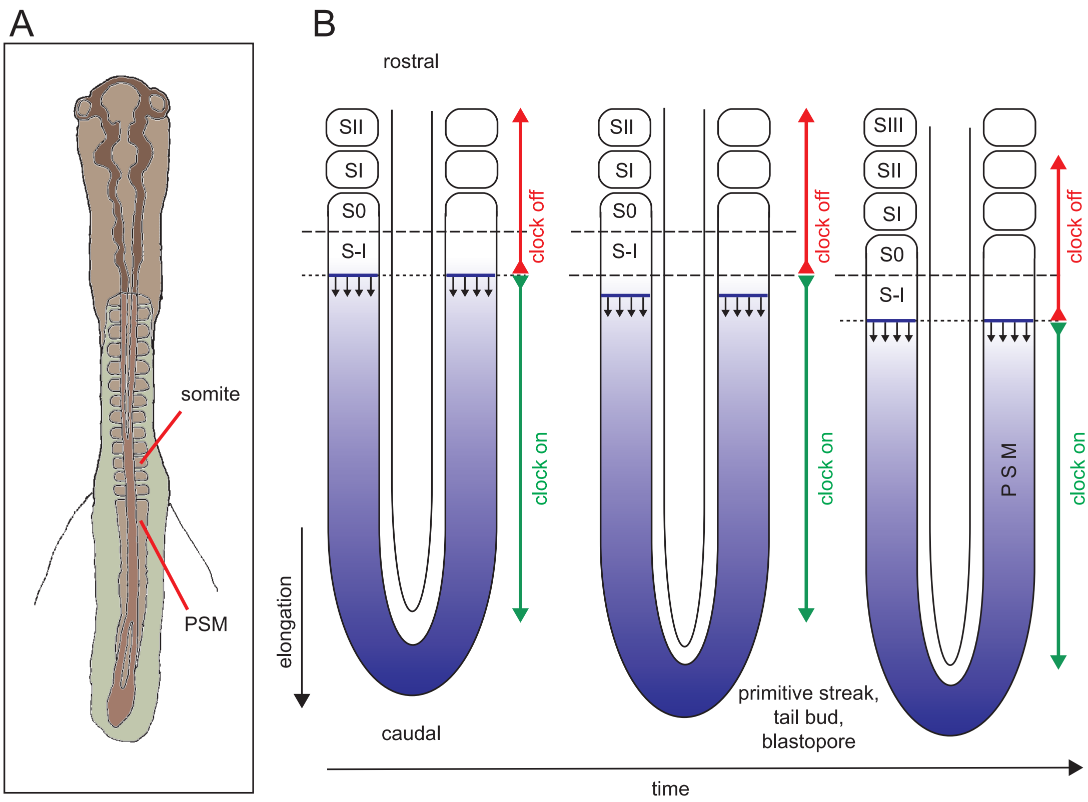


Figure 1.2: Somite formation in the vertebrate embryo. **(A)** Schematic illustration of a chicken embryo. **(B)** While the embryo elongates at the tail bud, pairs of somites regularly pinch off synchronously from the anterior tip of the presomitic mesoderm (PSM) in an anterior-to-posterior sequence. The morphogenic Wnt3a/Fgf gradient (blue) moves in caudal direction through the PSM. It acts as a determination front (blue line) and defines in combination with the intracellular clock the position of the border between prospective somites. Aulehla and Herrmann (2004) proposed that the morphogenic gradient drives the clock ("clock on") and, if cells are below a certain threshold level of the morphogen the clock cannot enter a new cycle ("clock off", dashed line) and a new somite boundary is defined. Thus the size of a somite is given by the distance passed by the determination front during one oscillation of the segmentation clock. Somites are denoted by SI, SII, SIII, etc. where the most recently formed somite is SI. Prospective somites are denoted by S0, S-I, S-II etc. (adapted from Pourquié and Tam, 2001; Aulehla and Herrmann, 2004; Dequéant and Pourquié, 2008)

Table 1.2: Clusters of some of the cyclic genes identified by Dequéant et al. (2006) using microarrays. Genes of the Notch-FGF cluster show a mutually exclusive activation compared to those of the Wnt cluster.

Notch cluster	*Hes1*	*Lfng*	*Nrarp*	*Nkd1*	*Hes5*	*Hey1*	*Bcl9l*	
Fgf cluster	*Spry2*	*Efna1*	*Hspg2*	*Egr1*	*Dusp6*	*Bcl2l11*	*Shp2*	
Wnt cluster	*Axin2*	*Dact1*	*Myc*	*Has2*	*Dkk1*	*Sp5*	*Tnfrsf19*	*Phlda1*

and Zeeman (1976). They postulated the existence of a "clock" and a maturation wave called the "wavefront". In that model the clock is assumed to be an intracellular oscillator that is phase-linked throughout the embryo and the wavefront is a front of rapid cell change moving slowly down the long axis of the embryo. When cells are in their permissive phase of the oscillator while passing the wavefront, they undergo a rapid alteration in locomotory and/or adhesive properties. According to their anterior-posterior body position the wavefront hits the cells at successively later time points.

Molecular evidence supporting the clock-and-wavefront model has been found. Aulehla and Herrmann (2004) proposed a model that takes into account a morphogen gradient established by the signaling molecules Wnt3a and Fgf8. This gradient is placed along the PSM. Both *Wnt3a* and *Fgf8* are expressed in the tail bud and, while the embryo grows at the tail bud in caudal direction, the concentration of these molecules decays during further elongation of the embryo, since the expression of these signaling molecules is restricted to the tail bud area. The molecular clock is supposed to be established by genes of the Wnt, Notch and FGF signaling pathways and their target genes. The segmentation process is illustrated in Fig. 1.2B. While the embryo elongates at the tail bud, pairs of somites regularly pinch off synchronously from the anterior tip of the presomitic mesoderm (PSM) in an anterior-to-posterior sequence. The morphogenic Wnt3a/Fgf gradient moves in caudal direction through the PSM. It acts as a determination front and defines in combination with the intracellular clock the position of the border between prospective somites. Aulehla and Herrmann (2004) proposed that the morphogenic gradient drives the clock ("clock on") and, if cells are below a certain threshold level of the morphogen, the clock stops ("clock off"). When the determination front passes cells that are in the permissive phase of the segmentation clock a new somite border is defined. Thus the size of a somite is given by the distance passed by the determination front during one oscillation of the segmentation clock.

1.2.2 Cell-cell Communication and Signal Transduction

For development and survival, cells must be able to react to changing environmental conditions. Therefore, during evolution, different mechanisms for the perception, intracellular transduction and interpretation of signals coming from outside the cell has evolved. Cells react with an appropriate response to the signal, like e.g., adaptation of the metabolism to compensate external stress. In addition to intracellular adaptation, cell-cell comunication is also essential for the development of multicellular organisms. Different mechanisms for the transmission of signals from one cell to another are known (Gilbert, 2003). In juxtacrine interactions cell membrane proteins on one cell surface interact with receptor proteins on an adjacent cell surface. This can only happen, when cells are situated next to each other. An example for a juxtacrine interaction is the interaction between the Notch receptor and the Delta ligand (cf. Fig. 1.3). Another mechanism of short distance cell-cell communication is the paracrine interaction, where signaling proteins (also called paracrine factors or growth and differentiation factors, GDFs) synthesized by one cell diffuse over a small distance to induce changes in nearby cells. This happens for instance when Wnt and FGF signaling is activated by their respective extracellular signaling molecules. A third mechanism of cell-cell communication is based on endocrine factors (hormones) that are secreted into the blood and travel to places far away from their production site to exert their effects. Notch, Wnt and FGF signaling are signal transduction pathways belonging to the first and second interaction mechanism, respectively. These pathways and their function during somitogenesis will be discussed in the following in more detail.

1.2.2.1 Notch Signaling

Notch signaling, triggered by a juxtacrine interaction, transmits signals between cells that are in direct contact with each other. Core components of the Notch signaling apparatus are (1) a Delta-type ligand, (2) a Notch-type receptor and (3) a transcription factor of the CBF1/Su(H)/LAG1 (CSL) family. The canonical Notch signaling pathway is depicted in Fig. 1.3. Proteins of the Delta- and Notch-type are single-pass transmembrane proteins carrying repeats of the epidermal growth factor (EGF) motif extracellularly (Rebay et al., 1991). A characteristic of EGF repeats is that they mediate direct contact between a ligand and a receptor (Rebay et al., 1991). When complexed to a Delta-type ligand (in mammals these are the Delta-like ligands DLL1, DLL2, and DLL3, and the Jagged ligands JAG1 and JAG2), Notch undergoes a conformational change. Once this has taken place, its cytoplasmatic domain can be cleaved by the protease Presenilin1, a member of the complex γ-secretase and the Notch intracellular domain (N_{ICD}) is released. The peptide translocates into the nucleus and binds to a dormant transcription factor of the CSL family thereby replacing a co-repressor

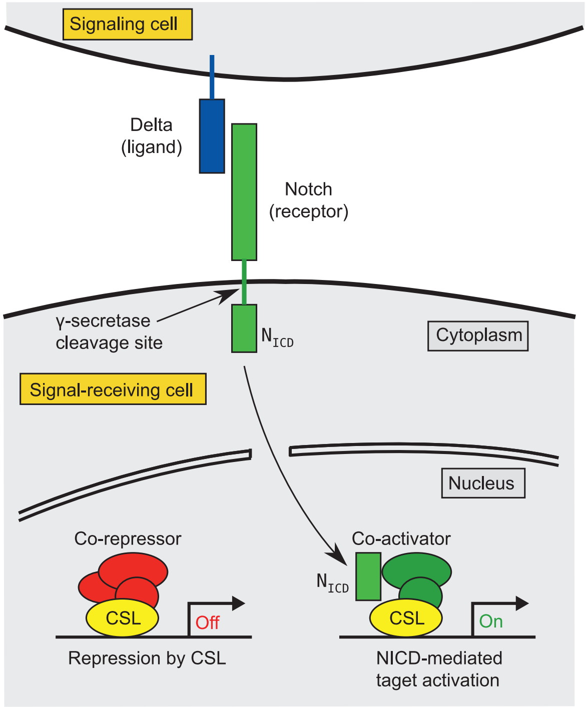


Figure 1.3: The canonical Notch signaling pathway. Key-players of the pathway are a Delta-type ligand, the Notch receptor and a transcription factor of the CSL family. When complexed with a Delta-type ligand, a part of the cytoplasmatic domain of the Notch protein is cleaved off. This Notch intracellular domain (N_{ICD}) translocates into the nucleus and acts as a co-activator of the transcription factor CSL activating target-genes of the Notch signaling. This activation happens by a replacement of the CSL co-repressor complex by a co-activator complex (adapted from Lai, 2004).

and activating the transcription factor (Lai, 2004; Gilbert, 2003). This implies that N_{ICD} is usually necessary for the activation of Notch target genes, but it is by far not sufficient to fulfill this task. Indeed, each of the Notch targets is not always activated when Notch signaling is active. The expression of a specific Notch target gene is co-regulated by other transcription factors and/or signaling pathways (Bray and Furriols, 2001). This complex mechanism of gene regulation allows for the activation of specific genes that are appropriate for different developmental settings. Thus a major biological role of Notch signaling is to control the developmental fates of cells and the regulation of pattern formation. Whether a cell predominantly expresses the ligand or the receptor is of high significance in this context.

1.2.2.2 Wnt Signaling

Wnt signaling is a paracrine interaction and it acts in numerous cellular processes including cell proliferation, survival, and differentiation. It thus has a significant impact on development and disease (Logan and Nusse, 2004; Moon et al., 2004). A simple outline of the current model of the canonical Wnt signaling pathway is shown in Fig. 1.4. A central player of this pathway is the protein β-catenin. It functions as a co-activator of genes regulated by the DNA-binding proteins of the lymphoid enhancer-binding factor 1 (Lef) family or the T cell-specific transcription factor (Tcf) family, with which β-catenin can form heterodimers. Free cytoplasmatic β-catenin has a high turnover-rate. When Wnt signaling is turned off, β-catenin is continuously phosphorylated by the active glycogen synthase kinase 3 β (GSK3β),

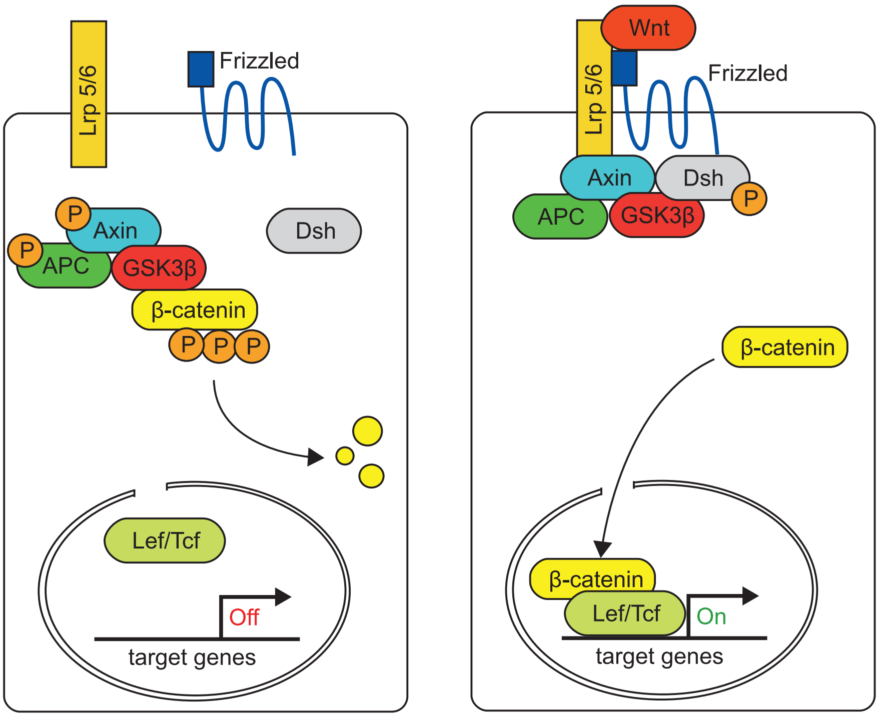


Figure 1.4: The canonical Wnt/β-catenin signaling pathway. In cells that are not exposed to the extracellular signaling molecule Wnt (left panel), the scaffold proteins Axin and APC can recruit GSK3β for the continuous phosphorylation of β-catenin, that becomes subsequently poly-ubiquitinated and degraded by the proteasome. Thereby the concentration of β-catenin remains low and genes regulated by β-catenin as a co-activator will not be transcribed. When cells are exposed to Wnt, the Frizzled receptor supported by the Lrp5/6 receptor can bind this glycoprotein. The perception of the extracellular signal activates Dsh and recruits the destruction complex (Axin/APC/GSK3β) to the membrane, where Axin is subsequently dephosphorylated and committed to destruction. This results in a decreased degradation of β-catenin by continuous Axin-dependent phosphorylation mediated by the Axin/APC/GSK3β complex. Thereby, unphosphorylated β-catenin accumulates in the cytoplasm and nucleus, and finally interacts with Tcf/Lef to control transcription of target genes (adapted from Logan and Nusse, 2004; Reya and Clevers, 2005; Cadigan and Liu, 2006).

which is part of a large destruction complex formed by the scaffold proteins Axin and adenomatous polyposis coli (APC). Phosphorylated β-catenin is a substrate for poly-ubiquitination and eventually proteasome-mediated degradation. Free GSK3β has a very low phosphorylation activity for β-catenin, but complexed with Axin and APC its phosphorylation activity for β-catenin increases tremendously (Dajani et al., 2003). Wnt signaling is activated by secreted Wnt ligands, cystein-rich glycoproteins that can interact with members of the frizzled (*Fz*) family, seven-transmembrane receptor proteins, and probably also with the single-pass transmembrane protein low density lipoprotein (LDL) receptor-related proteins 5 and 6 (*Lrp5* and *Lrp6*). Both, *Fz* and *Lrp5/6* act as receptors of the recipient cell and probably interact with each other when binding the Wnt ligand. Binding of Wnt to Fz, which is the primary receptor for Wnts (Bhanot et al., 1996), leads to the activation (phosphorylation) of the intracellular phosphoprotein Dishevelled (DSH or DVL) that probably recruits Axin and the destruction complex to the plasma membrane. This probably results in a dephosphorylation and degradation of Axin (Tolwinski and Wieschaus, 2004) that is presumably supported by phosphorylation thereby inhibition of GSK3β by active protein kinase B (PKB, Akt) (Naito et al., 2005). As a consequence the cytoplasmatic and nuclear level of β-catenin increases. Finally, by interaction with Lef/Tcf β-catenin activates the transcription of target genes, and thus Wnt signaling is switched on (Logan and Nusse, 2004).

1.2.2.3 FGF Signaling

Signaling mediated by fibroblast growth factors (Fgf) has been demonstrated to play a major role in embryonic, fetal and postnatal vertebrate development (Goldfarb, 1996; Martin, 1998; Böttcher and Niehrs, 2005). Fgf molecules are secreted proteins belonging to the paracrine signaling factors (Gilbert, 2003; Thisse and Thisse, 2005). Fgf molecules can bind to specific Fgf receptors (Fgfr), which are located in the cell membrane and are members of a large group of receptor tyrosine kinases. In human and mouse twenty-two different *Fgf* genes are known (Ornitz and Itoh, 2001; Itoh and Ornitz, 2004, 2008) that signal by activating a smaller family of cell surface receptors encoded by four distinct genes (Fgfr1–4), which can produce numerous Fgfr isoforms through alternative splicing (Johnson and Williams, 1993; Schlessinger, 2000). Fgfr receptors are single-pass transmembrane proteins with cytosolic tyrosine kinase activity.

FGF signaling (Fig. 1.5) is induced by binding of an Fgf ligand to an Fgf receptor and the subsequent assembly of receptor homo- or heterodimers (Ullrich and Schlessinger, 1990; Bellot et al., 1991) resulting in autophosphorylation of multiple tyrosine residues of the Fgfr receptor (Goldfarb, 1996; Mohammadi et al., 1996). Furthermore, it has been discovered that for the Fgf/Fgfr interaction heparin or heparin sulfate proteoglycans (HSPG) are required, which stabilize the formation of the receptor dimer (Yayon et al., 1991; Schlessinger et al., 2000).

Signaling complexes are recruited by the active Fgf receptor complex resulting in multiple phosphorylation events. One of these events is the activation of the Ras/mitogen activated protein kinase (MAPK) cascade which activates, amongst others, Erk that in turn regulates the activity of downstream kinases or transcription factors. The adaptor protein Frs2α has been shown to link Fgfr activation to the Ras/MAPK cascade. The PTB (phosphotyrosine binding) domain of Frs2α interacts with Fgfr (Ong et al., 2000), following the tyrosine phos-

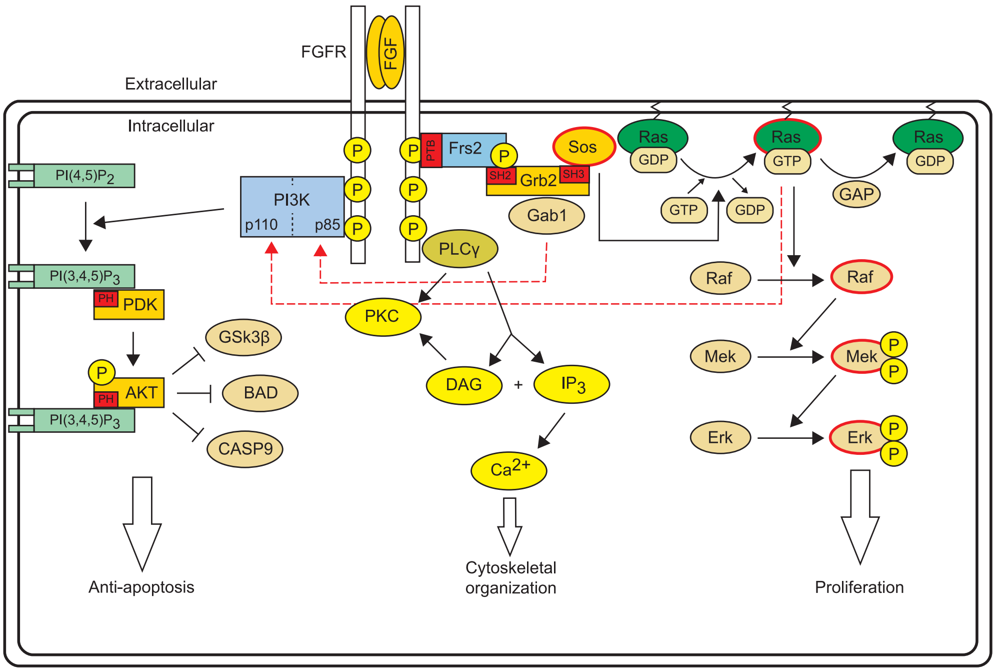


Figure 1.5: Overview of FGF signaling. Activated Fgfr can stimulate multiple pathways. It can result in an activation of PI3K/Pdk/Akt, PLCβ/PKC, or the Ras/MAPK cascade. A detailed description of Fgf activated pathways is given in the main text. (Alberts et al., 2008; Groth and Lardelli, 2002; Böttcher and Niehrs, 2005; Dailey et al., 2005).

phorylation of Frs2α by active Fgfr. Via its SH2 (Src homology 2) domain the adaptor protein Grb2 can bind to phosphorylated Frs2α and, in addition, recruit Sos to the plasma membrane that is linked with its proline-rich sequence motif to the SH3 (Src homology 3) domain of Grb2. Sos acts as a guanine nucleotide exchange factor (GEF) for the membrane associated GTPase Ras. Sos mediated exchange of GDP by GTP turns Ras into its active form Ras/GTP. Ras/GTP in turn activates a cascade of MAP kinases, in which active Raf phosphorylates and activates Mek that in turn activates Erk by phosphorylation. The deactivation of the MAP kinases is facilitated by phosphatases. The deactivation of Ras is mediated by a GTPase activating protein (GAP) that stimulates the GTPase activity of Ras whereby inactive Ras/GDP is formed.

Another pathway that concomitantly can get activated by FGF signaling is the PI3-kinase/Akt pathway (Fig. 1.5). Three different routes are described by which the PI3-kinase/Akt pathway can get activated (Böttcher and Niehrs, 2005). First, phosphatidylinositol 3-kinase

(PI3-kinase) directly binds to the active Fgfr receptor. Second, via Frs2/Grb2 the Grb2-associated binder-1 (Gab1) docking protein is bound and gets tyrosine-phosphorylated, resulting in the recruitment and activation of PI3-kinase (Hadari et al., 2001; Ong et al., 2001). Third, Ras/GTP can recruit the catalytic subunit p110 of PI3-kinase to the plasma membrane and activate it (Rodriguez-Viciana et al., 1994; Pacold et al., 2000). When activated, PI3-kinase phosphorylates phosphatidylinositol 4,5-bisphosphate (PI(4,5)P_2) at the 3 position of the inositol ring resulting in phosphatidylinositol 3,4,5-trisphosphate (PI(3,4,5)P_3). PI(3,4,5)P_3 is associated to the plasma membrane and can be bound by proteins having a pleckstrin homology (PH) domain. By this a kind of interaction phosphoinosite-dependent protein kinase 1 (PDK1) and Akt (also called protein kinase B, or PKB) are recruited to the plasma membrane. PDK1 phosphorylates and by this activates Akt. Active Akt acts as a kinase and performs phosphorylation of multiple target proteins. One of these targets is the glycogen synthase kinase-3β (GSK3β) that also has a pivotal role in Wnt signaling.

A third target that is activated by FGF signaling is the phospholipase Cγ (PLCγ). PLCγ activates the inositol phospholipid signaling pathway by cleavage of PI(4,5)P_2 into inositol 1,4,5-trisphosphate (IP_3) and diacylglycerol. Both molecules act as second messengers triggering an increase of the interacellular Ca^{2+} level and an activation of protein kinase C (PKC).

Several components of the FGF signaling are supposed to be relevant for somitogenesis and are included in the mathematical model that I have developed in the course of my thesis, presented in Section 2.2.3.

1.3 Computational Modeling of Biological Systems

Mathematical modeling and computer simulations can help to understand the internal nature and dynamics of complex systems such as biological systems and they can help to reveal links and relations that are not directly obvious. The development of a computer model for a given biological system involves several steps (Klipp et al., 2005, p. 9). At first one has to outline the problem that should be addressed by the model and formulate questions that should be answered by it. Later on, one has to collect all the data that is required for its implementation. Next, one has to decide about the model structure. This includes (1) the level of description, e.g., whether it deals with interacting molecules or cells, (2) the choice of a deterministic or stochastic approach, (3) the use of discrete or continues variables, and (4) the choice of a steady-state, temporal, or spatio-temporal description.

A very common way of modeling biological systems makes use of ordinary differential equations (ODEs). This approach is described in more detail in Section 1.3.1 where I introduce frequently used kinetic laws, which are also applied in Section 2.2 for the development of a mathematical model on somitogenesis.

Creating a fundamental knowledgebase on cellular reactions and their components is the first essential step in the development of computer models for cellular processes. This is already done by several projects resulting in large pathway databases that are described in Section 1.3.2 and that provide a valuable resource for modeling and modeling tools. Moreover, there are also databases on kinetic parameters or even detailed kinetic models of particular processes or pathways. Furthermore, I introduce most important standards for exchanging biological models and pathway data.

In Section 1.3.3 I present state-of-the-art tools for dynamic computational modeling. Computer tools allow the analysis of the dynamic behavior of the reaction networks given the model parameters. Very important features of such systems are for instance the estimation of model parameters from experimental data and the analysis of the behavior of the system with respect to changes of these parameters. I give an overview on the features of several tools.

Finally in this section I present published mathematical models describing different aspects of somitogenesis.

1.3.1 Mathematical Modeling of Biological Systems Using Ordinary Differential Equations

The compilation of mathematical models for biological systems requires knowledge about many system components (e.g., genes, enzymes, regulators, metabolites) as well as their different states (e.g., active, phosphorylated, methylated, etc.) and the interactions they participate in. Latter involves the stoichiometry of the reactants (substrates and products, i.e. the objects that are converted quantitatively) and the components which influence the reaction directly, but are not consumed or produced, i.e. they leave the reaction unchanged. This defines the structure (topology) of a model. Another information that is relevant to the development of continous models are the reaction kinetics. For each reaction of a model one has to know the kinetic law and its parameters, or make plausible assumptions for it.

1.3.1.1 Modeling of Biochemical Reactions

Kinetics of biochemical reactions can be described by the mass action law, which says that the reaction rate is proportional to the probability of the collision of the respective reactants (Guldberg and Waage, 1879).

For a reversible reaction of the form

$$n_1\mathrm{S}_1 + n_2\mathrm{S}_2 + \cdots + n_i\mathrm{S}_\mathrm{i} \underset{v_{\leftarrow}}{\overset{v_{\rightarrow}}{\rightleftharpoons}} m_1\mathrm{P}_1 + m_2\mathrm{P}_2 + \cdots + m_j\mathrm{P}_\mathrm{j} \tag{1.1}$$

with substrates S_i and products P_j the general mass action law reads

$$v = v_{\rightarrow} - v_{\leftarrow} = k_{\rightarrow} \prod_i [S_i]^{n_i} - k_{\leftarrow} \prod_j [P_j]^{m_j}, \tag{1.2}$$

where $v_{\rightarrow}$ and $v_{\leftarrow}$ are the respective reaction rates of the forward and backward reactions, $k_{\rightarrow}$ and $k_{\leftarrow}$ are their respective kinetic or rate constants, and $[S_i]$ and $[P_j]$ are the substrate and product concentrations with their respective molecularities n_i and m_j.

The concentration change of the substrates and products respectively is given by

$$\frac{d[S_i]}{dt} = n_i\, v \qquad \frac{d[P_j]}{dt} = -m_j\, v\,. \tag{1.3}$$

An important assumption for the classical deterministic kinetic modeling as it is described here, is that all reactants are homogeneously distributed.

Based on the mass action law, kinetics for several specific reaction mechanisms can be derived. For instance, for the irreversible enzymatic one-substrate reaction catalyzed by E that reads

$$\mathrm{E} + \mathrm{S} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{ES} \xrightarrow{k_2} \mathrm{E} + \mathrm{P} \tag{1.4}$$

Michaelis and Menten (1913) derived a kinetic law that was later extended by Briggs and Haldane (1925). It reads as follows:

$$v_{\mathrm{MM}} = \frac{V_{max}\,[S]}{[S] + K_m}\,. \tag{1.5}$$

This kinetic law shows a saturation behavior (Fig. 1.6A), with a maximum (V_{max}) that is proportional to the enzyme concentration ($V_{max} \propto E$). K_m is the substrate concentration for which the reaction rate is half maximal.

If a protein or enzyme has several binding sites instead of a single one, e.g., a protein complex that is composed of several subunits like a homotetramer, the binding of one ligand may change the binding affinity to further ligands. This phenomenon is called cooperativity. It has already been described in 1910 by Hill for the binding of oxygen to hemoglobin. A kinetic law describing this behavior is given by the Hill equation. Let us assume we have an enzyme E_2 with two binding sites for the substrate S and the following reaction

$$\mathrm{E}_2 + 2\mathrm{S} \longrightarrow \mathrm{E}_2\mathrm{S}_2 \tag{1.6}$$

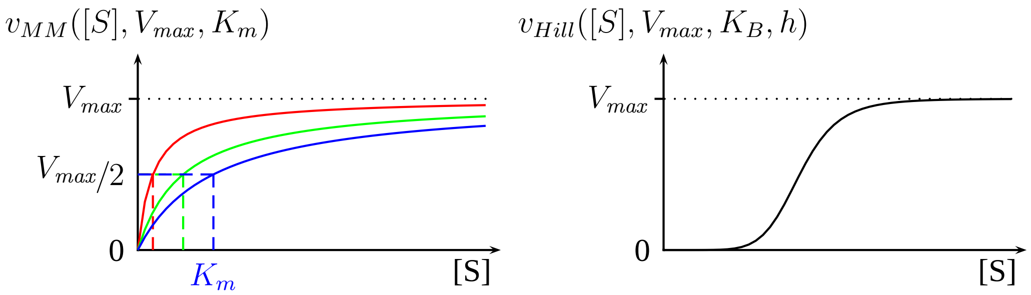
and the binding constant is defined by

$$K_B = \frac{[E_2 S_2]}{[E_2] \cdot [S]^2} \tag{1.7}$$

then the Hill equation reads

$$v_{Hill} = V_{max} \frac{K_B [S]^h}{1 + K_B [S]^h}, \tag{1.8}$$

where the quantity h is denoted the Hill coefficient. An example of the Hill kinetic is plotted in Fig. 1.6B.



(A) Michaelis Menten function

(B) Hill function (with $h > 1$)

Figure 1.6: Examples of some standard kinetic laws for biochemical reactions. In **(A)** the effect of different K_m values is illustrated. **(B)** Example of a Hill kinetic.

For a set of reactions the concentration change of a single component is given by the sum of in- and out-fluxes as follows

$$\frac{\mathrm{d}[S_i]}{\mathrm{d}t} = \sum_{j=1}^{r} n_{ij} v_j \qquad \text{with} \quad i = 1, \ldots, m \tag{1.9}$$

(Glansdorff and Prigogine, 1971). In this ordinary differential equation (ODE) system $[S_i]$ denotes the stoichiometric coefficient of the ith component, v_j is the reaction rate of the jth reaction, and n_{ij} is the concentration of the ith component in the jth reaction. The mathematical model is described by systems equations or balance equations (Equation 1.9). To do time course simulations this ODE system can be solved by the use of a numerical integrator or, if it is simple enough, also analytically.

Besides biochemical conversion reactions, association and dissociation processes are crucial for modeling of cellular interaction networks. Association and dissociation usually describe a reversible process of two or several components that form a single complex. The rate of dissociation can be described by the mass action law and its dissociation constant K_D, a specific equilibrium constant that measures propensity of a complex to dissociate. The inverse value of K_D is known as the association or affinity constant. For example, for the following reaction

$$\mathrm{A_nB_m} \xrightarrow{\qquad} n\mathrm{A} + m\mathrm{B} \tag{1.10}$$

K_D is defined as

$$K_D = \frac{[\mathrm{A}]^n \cdot [\mathrm{B}]^m}{[\mathrm{A_nB_m}]}\,. \tag{1.11}$$

1.3.1.2 Modeling of Gene Expression

Similarly as for metabolic reactions, gene regulation can also be described by coupled differential equations. The change in the level of each gene's mRNA can be introduced by two different terms, a positive term for transcription (mRNA synthesis) and a negative term for mRNA degradation. The expression of a gene, i.e. its transcription, depends on one or several other components known as transcription factors. For this the rate law for the mRNA synthesis of a gene—or, if transcription is neglected, the corresponding protein synthesis—depends on the concentration of the respective transcription factors.

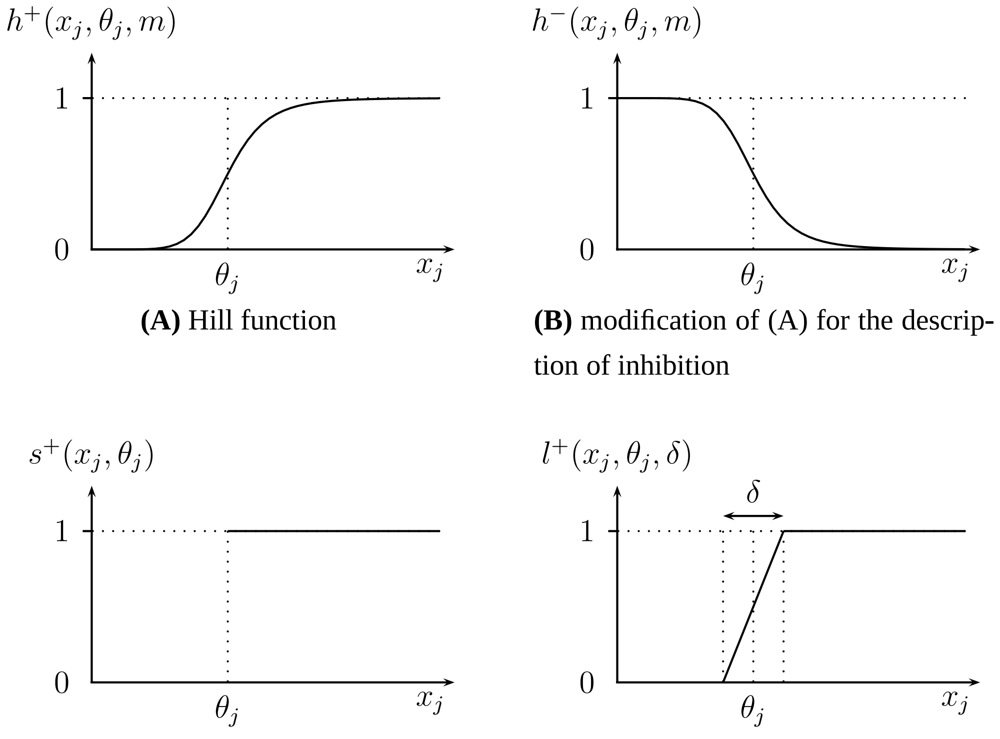
A kinetic law that is often found in literature for the description of gene regulation by a modifier (co-activator) is the Hill kinetic in a slightly modified form compared to Eq. 1.8:

$$h^+(x_j, \theta_{ij}, m) = \frac{x_j^m}{x_j^m + \theta_{ij}^m} \tag{1.12}$$

where $\theta_{ij} > 0$ is the threshold value for the influence of transcription factor j on the expression of gene i, and $m > 0$ a steepness parameter. The function ranges from 0 to 1, and increases monotonically as $x_j \rightarrow \infty$ (Fig. 1.7A). In order to express a repression in which x_j is an inhibitor one can use $h^-(x_j, \theta_{ij}, m) = 1 - h^+(x_j, \theta_{ij}, m)$ ((Fig. 1.7B); de Jong (2002)). For $m > 1$, the Hill kinetic has a sigmoid shape that is in agreement with experimental evidence (Yagil and Yagil, 1971). As an alternative to a Hill kinetic, gene expression can, e.g., also be described by non-continuous functions such as a step function (Fig. 1.7C) or a logoid function (Fig. 1.7D).

A general description for the kinetic modeling of gene regulation has been introduced by Schilstra and Bolouri (2002) and Schilstra and Nehaniv (2008). They give a logic semantic for the description that takes also into account inhibition and activation as well as effects like cooperativity and competition. A simplified rate law that can also be derived from the general description given by Schilstra and Bolouri is introduced by Mendes et al. (2003) and reads as follows

$$v_i = V_i \cdot \prod_j \left(\frac{K_{i_j}^{n_j}}{[I_j]^{n_j} + K_{i_j}^{n_j}} \right) \times \prod_k \left(\frac{[A_k]^{n_k}}{[A_k]^{n_k} + K_{a_k}^{n_k}} \right) \tag{1.13}$$



(A) Hill function

(B) modification of (A) for the description of inhibition

(C) step function

(D) logoid function

Figure 1.7: Examples of kinetic laws often used for gene regulatory processes. **(A)** Hill function h^+, **(B)** a modification of (A) that can be used for the descripton of inhibition $h^- = 1 - h^+$, **(C)** step function s^+, **(D)** logoid function l^+, (de Jong, 2002).

or a modification of this

$$v_i = V_i \cdot \prod_j \left(\frac{K_{i_j}^{n_j}}{[I_j]^{n_j} + K_{i_j}^{n_j}} \right) \times \prod_k \left(1 + \frac{[A_k]^{n_k}}{[A_k]^{n_k} + K_{a_k}^{n_k}} \right) . \tag{1.14}$$

In Eq. 1.14 the inhibitors I_j and activators A_k act independently of each other. V_i is a basal rate of transcription, i.e. when there is no action of inhibitors or activators. The constants K_{i_j} and K_{a_k} indicate concentrations at which the effect of the respective inhibitor or activator is half of its saturating value. The Hill coefficients n_j and n_k regulate the sigmoidicity of the curve. This kinetic description of gene regulation is used in the mathematical model on somitogenesis described in Section 2.2.

1.3.2 Data Resources for Systems Biology

The development of mathematical models of cellular systems requires a lot of information on different aspects of the system. Data typically arises from several levels of cellular information quantified by different functional genomics technologies such as DNA, RNA or protein sequence data, gene expression data from array experiments, abundance data of proteins and metabolites from diverse experimental techniques (e.g., mass spectrometry, 2D-gels, blots), information on protein-protein interactions or protein modifications, or kinetics of enzyme activities or binding affinities, among others. The most important resource for such information is the scientific literature and human expertise agglomerated in public databases. In particular for the development of mathematical models, standardized resources that provide their data in a computational amenable and reusable manner are a preferable resource. Tab. 1.3 gives a brief list of some important databases. A large compilation of relevant database resources is given in Galperin (2008). Moreover, the journal Nucleic Acids Research offers a yearly database issue in January, providing a broad overview of diverse databases.

1.3.2.1 Pathway and interaction databases

Pathway databases[2] are particularly interesting for modeling approaches since they offer a straightforward way of building network topologies by the annotated reaction systems. These databases provide integrated representations of functional knowledge of the different components of a biological system and constitute a foundation for the topology of mathematical models. The databases KEGG (Kanehisa and Goto, 2000; Kanehisa et al., 2008), Reactome (Joshi-Tope et al., 2005; Vastrik et al., 2007), and BioCyc (Karp et al., 2005) contain metabolic reactions and several signal transduction pathways. KEGG (Kyoto Encyclopedia of Genes and Genomes) is a reference knowledgebase offering information about genes and proteins, biochemical compounds, reactions, and pathways. It provides 240 reference pathways[3] that are linked to genes and reactions of multiple eukaryotes and many microorganisms. It can be accessed via the Web, FTP, and Web services. Reactome is managed as a collaboration of the Cold Spring Harbor Laboratory, the European Bioinformatics Institute (EBI), and the Gene Ontology Consortium. It uses a very precise specification (ontology) of components and interactions that comprises details on stoichiometry, localisation, references to external databases, etc. This covers also processes like complex formation events or translocations of molecules. A further pathway database with a similar scope is BioCyc that covers pathway data on *Escherichia coli* (EcoCyc), and predicted metabolic pathways of other microorganisms (MetaCyc), and human (HumanCyc). Databases with a specific focus on signaling events are BioCarta, Spike (Elkon et al., 2008), TRANSPATH (Schacherer et al., 2001), STKE, NetPath, and the Pathway Interaction Database (PID). An inherent aspect of the pathway concept is protein-protein interaction subject of the databases IntAct (Hermjakob et al., 2004; Kerrien et al., 2007) or DIP (Xenarios et al., 2000). Gene regula-

[2] Pathguide - the pathway resource list: `http://www.pathguide.org/`

[3] KEGG Release 48.0, Oct. 2008

Table 1.3: Databases useful for modeling of cellular systems.

Database	URL
Pathway databases	
KEGG	`http://www.genome.jp/kegg/`
Reactome	`http://www.reactome.org/`
BioCyc	`http://biocyc.org/`
EcoCyc	`http://ecocyc.org/`
MetaCyc	`http://metacyc.org/`
HumanCyc	`http://humancyc.org/`
BioCarta	`http://www.biocarta.com/`
Spike	`http://www.cs.tau.ac.il/ spike/`
TRANSPATH	`http://www.biobase.de/`
STKE	`http://stke.sciencemag.org/`
NetPath	`http://www.netpath.org/`
PID	`http://pid.nci.nih.gov/`
Protein interaction databases	
IntAct	`http://www.ebi.ac.uk/intact/`
DIP	`http://dip.doe-mbi.ucla.edu/`
Databases on gene regulation	
RegulonDB	`http://regulondb.ccg.unam.mx/`
TRED	`http://rulai.cshl.edu/TRED/`
TRANSFAC	`http://www.biobase.de/`
Databases on kinetic parameters	
BRENDA	`http://www.brenda-enzymes.info/`
SABIO-RK	`http://sabio.villa-bosch.de/`
Model databases	
JWS	`http://jjj.biochem.sun.ac.za/`
BioModels	`http://biomodels.org/`

tion processes and gene regulatory networks are not yet covered in such detail like metabolic processes or signaling. However, there are databases that store information on transcription factor binding sites such as RegulonDB (Salgado et al., 2006), TRED (Zhao et al., 2005), and TRANSFAC (Wingender et al., 2000; Matys et al., 2006). The lack of uniform data models and data access methods of the existing almost 224 interactions and pathway databases make data integration very difficult (Cary et al., 2005). Tab. 1.4 illustrates the overlap of several of these pathway resources in human.

Table 1.4: Numbers of overlapping reactions/interactions from different pathway databases that can be mapped to each other in respect of identical substrates and products (ConsensusPathDB, Oct. 2008; Kamburov et al., 2008).

	Reactome	Kegg	Humancyc	Pid	Biocarta	Netpath	Intact	Dip	Mint	Hprd	Biogrid	Spike
Reactome	4246	261	122	109	81	34	98	32	52	312	208	126
Kegg	261	1658	213	0	4	0	0	0	0	0	0	0
Humancyc	122	213	1322	0	2	0	1	2	2	7	3	2
Pid	109	0	0	3741	285	100	71	48	78	352	249	202
Biocarta	81	4	2	285	2221	69	52	36	44	145	115	173
Netpath	34	0	0	100	69	1915	58	34	124	819	508	235
Intact	98	0	1	71	52	58	6995	312	2816	3285	1621	4330
Dip	32	0	2	48	36	34	312	1216	393	823	638	443
Mint	52	0	2	78	44	124	2816	393	13176	7446	4545	5939
Hprd	312	0	7	352	145	819	3285	823	7446	37952	18721	11854
Biogrid	208	0	3	249	115	508	1621	638	4545	18721	28206	10738
Spike	126	0	2	202	173	235	4330	443	5939	11854	10738	22230

Besides topological information about cellular reaction networks, also kinetic data, like kinetic laws and kinetic constants, are of particular interest for the generation of mathematical models. Two databases that are concerned with such data are BRENDA (Schomburg et al., 2004) and SABIO-RK (Wittig et al., 2006).

Mathematical models of a biochemical reaction system have been made available to the scientific community in form of a publication often depicting a diagram of the reaction system or a list of the reaction equations, along with a mathematical description (e.g., as a differential equation system), and lists of kinetic parameters and concentrations of specific states. Recently, model databases have been setup, such as the BioModels database (Novère et al., 2006) or JWS (Olivier and Snoep, 2004). Both are free, centralised databases of curated, published, quantitative kinetic models of biochemical and cellular systems. For instance, the BioModels database currently provides 87 curated and 40 non-curated models.

1.3.3 Software Applications for Modeling and Simulation

The computation of time courses of a biochemical reaction system based on a given pathway structure and its kinetic scheme, that is required for simulations, has already been discussed

1970 by Garfinkel et al. It arises from fundamental research on biochemical reaction kinetics (e.g. Michaelis and Menten, 1913). The first simulation of a biochemical system (the peroxidase reaction) was carried out by Chance (1943), who used a mechanical differential analyzer to solve mathematical equations.

Table 1.5: Modeling tools frequently used in systems biology.

Application	URL
Gepasi	`http://www.gepasi.org/`
COPASI	`http://www.copasi.org/`
E-Cell	`http://www.e-cell.org/`
ProMoT/Diva	`http://www.mpi-magdeburg.mpg.de/projects/promot/`
Virtual Cell	`http://www.nrcam.uchc.edu/`
Systems Biology Workbench	`http://sys-bio.org/`
Cell Designer	`http://www.celldesigner.org/`
PyBioS	`http://pybios.molgen.mpg.de/`

During the past decades in the course of the computational revolution more and more software applications were developed that can be used for the description of the dynamic behavior of biological systems. In Tab. 1.5 several software applications are listed. A very comprehensive list of such software applications can be found at the SBML homepage[4].

Often general-purpose applications such as Mathematica (Wolfram Research) and Matlab (MathWorks) are used that are designed for the computation and visualization of any type of mathematical model. Although these software tools are very advanced, they have a steep learning curve, require a lot of mathematical background knowledge, and are not designed for the setup of biological models. This gave rise to the development of many other software applications that better meet the desired requirements.

One of the first applications designed for simulation of biochemical reaction systems is Gepasi that was developed in the beginning of the 1990ies. It is a stand-alone-application and comes up with a user-friendly interface for the simulation and analysis of biochemical systems (Mendes, 1993, 1997; Mendes and Kell, 1998). It provides time course and steady state simulation and the ability to explore the behavior of the model over a wide range of parameter values using a parameter scan that runs one simulation for each parameter combination. Gepasi can be used to characterize steady states using metabolic control analysis (MCA, Kacser and Burns, 1973; Heinrich and Rapoport, 1974) and linear stability analysis and is capable of doing parameter estimation with experimental data. The successor of Gepasi is COPASI that has similar but improved functions and some extensions (Hoops et al.,

[4] `http://sbml.org/SBML_Software_Guide/SBML_Software_Summary`

2006).

E-Cell is based on the modeling theory of the object-oriented Substance-Reactor Model (Tomita et al., 1999; Takahashi et al., 2003). Models are constructed with three object classes, Substance, Reactor, and System. Substances represent state-variables, Reactors describe operations on state-variables, and Systems represent logical or physical compartments. It provides different classes of standard Reactors (e.g., Michaelis-Menten formula). Time course calculation is done by the use of a simulation engine. Numerical integration is supported by first-order Euler or fourth-order Runge-Kutta method.

ProMoT/Diva consists of the modeling tool ProMoT and the simulation environment Diva (Ginkel et al., 2003). The workbench deals with modular models and can handle differential algebraic equation (DAE) systems. Modeling is supported by a graphical user interface and a modeling language. The modeling tool provides the possibility to use existing modeling entities out of knowledge-bases.

The Virtual Cell is a web-based client-server architecture with a central database of user models. It provides a formal framework for modeling biochemical, electro-physiological, and transport phenomena while considering the sub-cellular localization of the molecules that take part in them (Slepchenko et al., 2003).

The Systems Biology Workbench (SBW) provides a server that acts as a broker between different modeling and simulation tools (clients) via a common interface (Hucka et al., 2002). These clients (add-ons) cover graphical tools for model population, deterministic and stochastic simulators and analysis tools like the integration of MetaTool (Pfeiffer et al., 1999).

CellDesigner provides an advanced graphical model representation along with an easy to use user-interface and an integrated simulation engine (Funahashi et al., 2003). For the development of a model CellDesigner supports a rich set of graphical elements for the description of biochemical and gene-regulatory networks. Networks can be constructed from compartments, species, and reactions. CellDesigner comes with a large number of predefined shapes that can be used for different types of molecules, such as proteins, receptors, ion channels, small metabolites, etc. In CellDesigner it is also possible to indicate phosphorylations or other modifications. The program also provides several icons for special reaction types like catalysis, transport, inhibition, and activation.

All these software applications provide the ability to define a model step by step, e.g., by entering the reaction details as plain text or by the use of some graphical interfaces. Entering reaction details step by step is very important, but can become very cumbersome and error-prone, in particular for large biochemical reaction networks. In this context, the visualization of the reaction network of a model is also very important. Functions for the visualization of reaction networks are only provided by some of the above mentioned software applications. In particular none of them provide a flexible way for the automatic visualization of parts of a reaction network that is very important, e.g., for large models that are ofter very complex. Moreover, often one is interested in comparing simulation results directly with the underlying network structure, to understand the dynamic system behavior in the context of the reaction network. This is also a feature that is not provided by current software applications. To overcome these limitations it was necessary to develop a software application, namely PyBioS, that provides those features and could be used for the setup of a quite large model on somitogenesis that incorporates different signaling pathways and gene regulatory feedback mechanisms.

Since the diverse software applications provide different features, a well defined format for data exchange and documentation of the components and reactions of a model is pivotal. This demand resulted in the development of several data formats for pathway data and mathematical models. The BioPAX[5] format is a very general and expressive format and is designed for handling information on pathways and topologies of biochemical reaction networks. Other formats that are designed for the description of mathematical models of biochemical reaction systems are the Systems Biology Markup Language (SBML, Hucka et al., 2003, 2004) and CellML (Lloyd et al., 2004).

1.3.4 Mathematical Models of Somitogenesis

Mathematical modeling turns out to be significantly useful for the description and understanding of cellular processes and can be used for hypothesis testing and making experimentally testable predictions. There are already multiple mathematical models describing different cellular processes, such as metabolic pathways like glycolysis (e.g. Teusink et al., 2000; Hynne et al., 2001), signal transduction pathways, like MAP kinase signaling (e.g. Huang and Ferrell, 1996; Hatakeyama et al., 2003) or WNT signaling (Lee et al., 2003), gene-regulatory processes (e.g. Elowitz and Leibler, 2000), or the cell cycle (e.g. Goldbeter, 1991; Tyson et al., 1996; Novák et al., 1999).

There are also several mathematical models describing cellular processes of somitogenesis. For instance, Lewis (2003) has worked out a mathematical model for oscillation during somitogenesis in zebrafish. The model takes into account the *her1* gene and its corresponding protein that acts as a repressor for its own expression (Fig. 1.8A). The model is described by two time-dependent delay differential equations

[5] BioPAX: `http://www.biopax.org`

$$\frac{\mathrm{d}[p(t)]}{\mathrm{d}t} = a[m(t-T_p)] - b[p(t)], \tag{1.15}$$

$$\frac{\mathrm{d}[m(t)]}{\mathrm{d}t} = f([p(t-T_m)]) - c[m(t)] \tag{1.16}$$

where $[p]$ and $[m]$ denote the protein and mRNA concentrations, respectively, a is the translation rate, and b and c are the decay rates of the protein and mRNA, respectively. The mRNA synthesis is described by

$$f([p]) = \frac{k}{1+[p]^2/p_0^2}, \tag{1.17}$$

which describes the inhibitory effect of the protein that acts as a dimer on the mRNA transcription. $T = T_m + T_p$ is the time delay given due to transcription T_m and translation T_p. For sustained oscillation it is assumed that the lifetimes of the mRNA and protein are very short compared with the total decay time T. For sustained oscillation the peak of the protein concentration is shifted slightly behind that of the mRNA concentration.

Based on the work of Lewis (2003), Hirata et al. (2004) have worked out a model for the description of the oscillatory behavior of *Hes7* expression in mouse. They analysed the model in respect to the protein half live that turned out as a crucial parameter. Based on the simulation results of their model they could show that a half life of 20 min for the Hes7 protein provides a sustained oscillation, while an increase to 30 min results in a damped oscillation which is in accordance to experimental findings.

Zeiser et al. (2008) have converted the model of Hirata et al. (2004), which is described by two delay differential equations, into an ordinary differential equation system consisting of five different components (see Fig. 1.8B). It takes into account separate variables for both the mRNA and the protein. Furthermore, the cytosolic protein is first ubiquitinated before degradation. All reactions are described by linear kinetics except for the inhibition of the gene expression that is described by a Hill kinetic (cf. Section 1.3.1) and the ubiquitination that is described by a Michaelis-Menten kinetic (cf. Section 1.3.1). Zeiser et al. (2008) could mimic the qualitative results found by Hirata et al. (2004) according to the half life of the ubiqutinated protein without explicit specification of a time delay.

Although *her1* and *her7* in zebrafish can form a sustained oscillator, it is not sufficient to form a robust molecular clock for somitogenesis. Surprisingly, zebrafish embryos lacking both *her1* and *her7* or embryos injected with *her1* and *her7* morpholinos still form abnormal somites (Henry et al., 2002). This observation indicates that further components are involved in the molecular clock controlling somitogenesis. Orthologues of the zebrafish hairy and enhancer of split genes, whose expression oscillate during somitogenesis, have also been detected in amniotes (where they are called Hes genes), like chicken (*HES1*, *HAIRY2*, and

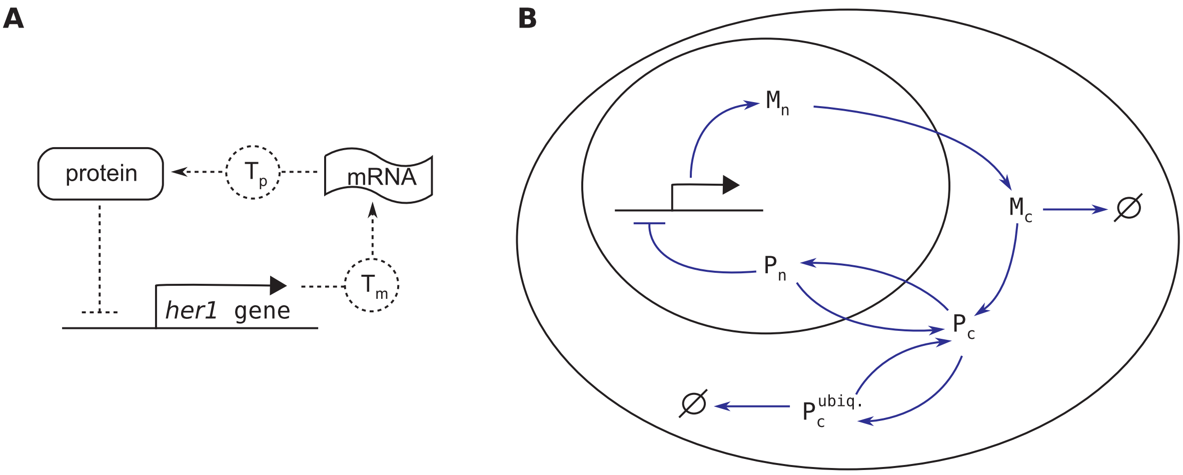


Figure 1.8: Models of molecular autoinhibitory circuits. (**A**) Model of *her1* autoinhibition proposed by Hirata et al. 2004. (**B**) Model of *Hes7* autoinhibition developed by Zeiser et al. (2008). The model considers compartmentalization where M_n and M_c are the mRNAs in the nucleus and cytosol, respectively, and P_n and P_c are the respective proteins each in the nucleus and cytosol. $P_c^{ubiq.}$ is the ubiquitinated protein targeted for degradation.

HEY2) and mouse (*Hes1*, *Hes7*, *Hes5*, and *Hey1*; Dequéant and Pourquié 2008). Moreover, cyclic expression of other Notch pathway genes were identified, as for example delta-like 1 (*Dll1*) and lunatic fringe (*Lfng*) in mouse. Today, it is well established that Notch signaling plays an important role in the clock mechanism. However, when Notch signaling is impaired or abolished, still somites can be formed, suggesting that additional factors must be involved, such as components of Wnt signaling and FGF signaling as proposed by Aulehla and Herrmann (2004). One component of the Wnt signaling that has been found to oscillate during somitogenesis is Axin2 (Aulehla et al., 2003). Based on different data that suggest an oscillation of Wnt signaling activity in the presomitic mesoderm (PSM), Aulehla and Herrmann (2004) proposed a negative feedback loop involving *Axin2* as a target of Wnt signaling and the subsequent destabilization of Axin2 protein that can form another molecular oscillator. Moreover, they outline that there is a tight link between Wnt and Notch signaling cascades in the oscillating part of the PSM and they suggest that the oscillations of Notch signaling activity are dependent on Wnt3a. Furthermore, *Fgf8* RNA was shown to form a gradient along the PSM and, by increasing the local concentration of Fgf8 protein in the PSM, the somite size was reduced, whereas the inhibition of FGF signaling results in larger somites (Dubrulle et al., 2001). These observations indicate the importance of Fgf8 in determining the position at which a segment boundary will form. In addition to the Fgf8 gradient that is formed in the PSM with a high concentration in the tail bud and a decrease in anterior direction, a second, parallel gradient formed by Wnt3a was also identified (Aulehla et al., 2003).

A first mathematical model of somitogenesis integrating Notch, Wnt and FGF signaling was developed by Goldbeter and Pourquié (2008). The model is set up by three separate models for FGF, Wnt and Notch signaling, respectively, each with independent oscillators. Goldbeter and Pourquié (2008) showed that coupling of the three oscillators can lead to synchronized oscillations in the three signaling pathways or to complex periodic behavior, depending on the relative periods of oscillations in the three pathways.

In the course of my thesis I have developed a comprehensive mathematical model of somitogenesis that includes additional components of Notch, Wnt and FGF signaling and assumes other cross-talks between the pathways. It is introduced in Section 2.2.

1.4 Experimental Techniques for Gene Expression Analysis

Besides the analysis of biological systems, modeling strategies can also be applied to biotechnological experimental techniques. One particular technique of high interest in molecular genetics is gene expression analysis. Today the genome sequence of several species is known. Among the first genomes to be sequenced have been those of some microorganisms like *Mycoplasma genitalium* (Fraser et al., 1995) or *Escherichia coli* (Blattner et al., 1997) and later also those of eukaryotes ranging from *Saccharomyces cerevisiae* (Goffeau et al., 1996) to mouse (Waterston et al., 2002), rat (Gibbs et al., 2004), and human (Lander et al., 2001; Venter et al., 2001). The availability of large scale sequence data gave rise to the development of new high-throughput technologies for transcriptome analysis, such as DNA arrays.

DNA array technologies are of particular interest, since the transcriptional state gives a snapshot of the gene expression state and thus an overview of the genes that might be active at a particular time point. Of course this information is biased by degradation rates, post-translational modifications, activations or inhibitions.

1.4.1 cDNA Array Technology

DNA array technology is nowadays frequently used in transcriptome analysis for the generation of genome-wide gene expression profiles[6]. DNA arrays benefit from the biochemical feature of hybridization. Hybridization describes the binding of two complementary strands of nucleic acids to each other via hydrogen bonds, where complementarity refers to the rule

[6] For a review see *The chipping forecast*, Nature Genetics, Vol. 21, Suppl. Issue 1, 1999; *The chipping forecast II*, Nature Genetics, vol 32 Suppl., 2002; or *The chipping forcast III*, Nature Genetics, Vol. 37, Suppl. (6s), 2005

of Watson and Crick. This rule says that adenine (A) can bind thymine (T) via two hydrogen bonds and cytosine (C) can bind guanine (G) via three hydrogene-bonds. For decades hybridization has been used in molecular biology for different techniques such as Southern blotting and Northen blotting. By these techniques DNA or RNA that was separated by gel electorphoresis is transfered to a filter membrane, and later exposed to a radioactively labeled oligonucleotide probe. DNA or RNA fragments which are complementary to the probe can such be identified.

DNA arrays are a massively parallel version of these techniques. Thousands of DNA samples are immobilized as spots within micrometers to each other on a surface (e.g. a nylon filter or a glass slide) and hybridized with a labeled sample. The immobilized samples usually have known sequences and are denoted as *probes*. The labeled sample that has to be identified is called the *target*. The target sample is usually derived from total mRNA of the cells which are under investigation. After digitalization of the hybridized array image, a numerical value, the signal intensity is assigned to each probe. It is assumed that this signal intensity is proportional to the number of molecules of the respective gene in the target sample, and hence changes in signal intensities can be interpreted as concentration changes. It should be pointed out that this is only valid as long as the intensity-concentration correlation is approximately linear. Nonlinearities might occur, for instance, by saturation effects or if the concentration falls below the detection limit of the DNA array.

By this extensive parallel expression analysis it is possible to study not one or a few genes at a time, as it is the case for Southern blots or Northern blots, but thousands of genes in parallel with a single experiment. Hence, DNA arrays are an ideal experimental platform for systems biology.

The first DNA array platform was the macroarray developed in the late 1980s (Poustka et al., 1986; Lehrach et al., 1990; Lennon and Lehrach, 1991). This technique employs PCR products of cDNA clones that are immobilized on nylon filter membranes and hybridized with radioactively labeled target material. The hybridization pattern is detected using a phosphor imager. cDNA macroarrays typically have a size of 8×12 cm^2 to 22×22 cm^2 and cover up to 80 000 different cDNA clones. Multiple studies employed this technique (Gress et al., 1992, 1996; Granjeaud et al., 1996; Nguyen et al., 1995; Dickmeis et al., 2001; Herwig et al., 2001; Kahlem et al., 2004).

cDNA microarrays are another DNA array platform. Here, cDNA is spotted on glass slides by a robot. The immobilized probes are hybridized by flourescently labeled target material. Microarray chips are small (1.8×1.8 cm^2) and allow the spotting of tens of thousands of different probes. cDNA microarrays are widely used in genome research (Schena et al., 1995, 1996; DeRisi et al., 1996, 1997; Spellman et al., 1998; Iyer et al., 1999; Bittner et al., 2000; Whitfield et al., 2002; Adjaye et al., 2005). A specific advantage of this technology is that the sample target and the control target can be labeled with different fluorochromes (Cy3 and Cy5 dyes; e.g. Amersham Pharmacia Biotech, Santa Clara, CA) and used for the same hybridization. Afterwards, two scanning procedures are performed for the different fluorochromes, respectively. This yields two images, one of the sample target and another of the control target.

A third platform is commercial oligonucleotide chips: Oligonucleotides are either spotted or *in situ* synthesized on slides. Latter method is applyed by the photolithographic procedure used for the production of Affymetix chips. These chips characterize a single gene by the use of a set of approximately 20 oligonucleotide probes of length 20–25 nucleotides (Lockhart et al., 1996; Wodicka et al., 1997; Lipshutz et al., 1999). These probes are denoted as perfect matches (PM), because they are perfectly complementary to parts of the mRNA of the respective gene. For the detection of nonspecific and background hybridization, mismatch (MM) oligonucleotides are synthesized that differ only in the central position 13 from the PM oligonucleotides. Chips are typically small (1.8×1.8 cm^2). The target sample is labeled with a single fluorochrome, so two chips are required to compare a sample and a control. Another commercial platform is that of Agilent. These chips are produced by an inkjet printing technology, known from printers, that has been adapted for the manufacturing. Agilent chips utilize 60mers as probes (Hughes et al., 2000, 2001).

1.4.2 Image Analysis

The mentioned DNA array platforms provide the experimental module of this gene expression analysis technique. The second part is data analysis that is done by the bioinformatics module. Output of the experimental platform is a digitized hybridization image. First step of the analysis pipeline is image analysis. In this step each probe spot of the scanned DNA array image is assigned a numerical value that represents the signal intensity. Essential for this is the correct identification of each spot center, and a correct quantification of the pixel neighborhood around the identified center of each spot. Since the signal intensities determined during image analysis are the input data to any further pre-processing steps and fold-change analysis or clustering analysis, the quality of image analysis is essential for any results that can be gaind by subsequent procedures.

Commonly, image analysis is a two-step procedure: In the first step, the grid finding, a grid is determined whose nodes describe the center postitions of the probe spots. In the second step, the quantification of signal intensities, a certain pixel area around the respective spot center is used to compute the signal intensity. For image analysis several commercial products are available, e.g., ImaGene (BioDiscovery), Genespotter (MicroDiscovery), GenePix (Axon), AIDA (Raytest), and Visual Grid (GPC Biotech). Moreover, academic groups have developed their own software, e.g., ScanAlyze (Stanford University), FA (Max Planck Institute for Molecular Genetics, Steinfath et al., 2001), and UCSF Spot (University of California, San Francisco, Jain et al., 2002). All these products differ in several points, e.g., the array platforms they are designed for, the degree of automation, and the usability. Very importanat points are of course the correct spot identification and quantification, that depends on the implemented algorithms, and manual settings required by the user, like clicking the corners of the spotted area. Hence, image analysis programs can be classified manual, semiautomated, and automated according to the degree of user interaction.

Grid Finding

Spots of the array are usually arranged in a rectangular grid. Due to experimental problems of the spotting procedure, the center of a spot is usually not exactly at its ideal grid position, e.g., sub-grids can be shifted against each other, spots can be distorted irregularly in each direction, and irregular spot shapes can make the spot identification more difficult. The purpose of grid finding is to assigne all spots to their corresponding grid position and to identify the correct center of each spot. The procedures comprise mostly geometric operations, like rotations and projections of the digitized image. In the first step of the grid finding the global borders of the originally reactangular grid are identified. During further steps smaller sub-grids are found, and finally the individual spot positions are identified. Common basic steps of the grid finding procedure are (1) a pre-processing of the pixel values, (2) the detection of the spotted area, and (3) the spot finding (Steinfath et al., 2001). The purpose of the first step is to amplify signal pixels, while reducing noise, e.g. by shifting a theoretical spot mask across the image and assign those pixels to grid nodes that show the highest correlation to the theoretical spot shape. Therefore, the theoretical spot shape should be similar to most of the spots, e.g., a two-dimensional gaussian distributed shape might be appropriated. The second step identifies the quadrilateral of the spotted area. For this step several of the above mentioned programs require user interaction by manual definition of the spotted area, e.g., by clicking the edges. They are semiautomated (e.g., Visual Grid). Fully automated programs provide an automatic corner detection (e.g., FA). In the third step of this procedure each node of the grid is detected and local maxima are identified that are the centers of the spots.

Quantification of Signal Intensities

Once the centers of the spots have been identified, a certain pixel area around each spot center is used to compute the signal intensity. Potential errors the quantification has to cope with are background noise due to unspecific binding, overshinig effects of spots that are next to each other, or irregular spot shapes. The quantification might be done in two different ways: Segmentation tries to separate the foreground pixels that belong to the spot, from its surrounding background pixels (Jain et al., 2002). Then, a spot intensity and potentially also a background value can be calculated from the respective areas. Another quantification method is the spot shape fitting that tries to fit a particular probability distribution, e.g., a two-dimensional Gaussian spot shape around the spot center. The signal intensity is computed as a weighted sum of the pixel intensities and the fitted density (Steinfath et al., 2001).

Databases of Expression Data

Microarray data provide a valuable resource in the interpretation of the transcriptome levels of genes. Large repositories store these data from multiple studies such as the Gene Expression Omnibus (GEO) (Edgar et al., 2002; Barrett et al., 2007) at NCBI and the ArrayExpress (Brazma et al., 2003; Parkinson et al., 2007) at EMBL-EBI. These databases provide free distribution and shared access to comprehensive gene expression datasets. Data include single and multiple channel microarray-based experiments measuring the abundance of mRNA, genomic DNA and protein molecules. Data from non-array-based high-throughput functional genomics and proteomics technologies are also archived, including SAGE, and mass spectrometry peptide profiling.

Reliability of Array-based Expression Data

The reliability of data produced by these experiments and their reproducibility are crucial for this research. To ensure both reliability and reproducibility a sophisticated experimental design is necessary. This includes for example the identification of error parameters that affect the hybridization data during the data generation process. Influences of systematic and statistical errors due to biotechnological methods (for example mRNA preparation, PCR, hybridization), as well as due to devices and array-media (for example robots, filters, glass-slides) and their effects on evaluation software and algorithms (image analysis, statistical tests, clustering algorithms) must be estimated. These sources of error are frequently discussed in the context of calibration and normalization of microarray data (e.g. Dudoit et al., 2002; Huber et al., 2002; Kepler et al., 2002; Schuchhardt et al., 2000).

In the course of my thesis I have developed a computer model for the simulation of cDNA macroarrays that takes into account several sources of error. It enables scientists to judge which parameters are critical and how the experimental design or data evaluation might be improved. The computer model is introduced in Section 2.3. Moreover, using this model, I performed simulations of DNA array hybridization experiments for the evaluation of critical parameters during subsequent image and data analysis.

1.5 Objectives

The objectives of my thesis are (1) the development of a modeling and simulation platform for biological systems, (2) the use of the modeling platform for the development of a molecular model of the mouse segmentation clock that plays a central role in somitogenesis, and (3) the application of modeling strategies on cDNA arrays.

Modeling and simulation platform. Computational models of biological systems are essential parts of systems biology. While the mathematical description of molecular reaction networks, e.g., by systems of ordinary differential equations (ODEs), is well established, the availability of advanced computational tools for the management and simulation of those systems that tend to be large and complex, is still subject to current research. Therefore, I envisage the development of a modeling and simulation platform for biological, in particular cellular and biochemical reaction systems. The computational tool shall be able to represent essential information of a molecular reaction system that is necessary for the construction of a mathematical model. The tool shall apply modern concepts of object-oriented programming that serves as a flexible structure for data representation and expandability. The system shall come with a user interface for the development of models and it shall serve as a model repository. Moreover, it shall be able to integrate data from public pathway databases in order to use those data for model development.

Development of a molecular model of somitogenesis. Somitogenesis is a fundamental developmental process taking place during vertebrate embryogenesis. There is evidence supporting a model of a morphogen gradient and a molecular clock responsible for the serial determination of somite formation. Aulehla and Herrmann (2004) have proposed a model of the molecular clockwork comprising Wnt, Notch and FGF signaling. It is assumed that the clock is driven by Wnt signaling downstream of Wnt3a. The morphogen gradient is established in mouse by Wnt3a and Fgf8, both are produced in the tail bud, but with a decay in the anterior PSM. The objective is to develop a mathematical model of the molecular clock and its connection to the morphogen gradient. The model shall be able to describe properties of somitogenesis, the arrest of the molecular clock below a certain Wnt3a concentration and provide evidence of experimentally observed oscillation of clock components.

Modeling of cDNA arrays. In engineering sciences modeling and simulation techniques have proven to be significantly helpful for the evaluation of technical processes. In a similar manner also modeling can be applied to laboratory methods of modern molecular biology. Gene expression analysis based on complex hybridization analysis have increased rapidly in about the last ten years. Although complex hybridization experiments are based on a data production pipeline that incorporates a significant amount of error parameters, the evaluation of these parameters has not been studied yet in sufficient detail. An objective of my thesis is to model cDNA hybridization experiments and to use the model for simulation and subsequent statistical evaluation of error parameters of the experimental data production pipeline.

2 Results

In the first part of this chapter I introduce the modeling and simulation system PyBioS that I have developed in the course of my thesis (Section 2.1). I describe the general concept and design of the PyBioS system, its user interface, its unique features and demonstrate its usability even for large biochemical reaction systems. In the second part of this chapter I show the application of modeling and simulation to an experimental laboratory method (DNA array experiment) and to cellular processes.

Furthermore, I have created a model for the developmental process somitogenesis. The model describes general features of the molecular segmentation clock known to take place in segmental pattern formation during embryonic development.

The presented work on experimental and biological systems illustrates the usability of modeling and simulation techniques for biology and shows its impact on current research in molecular biology.

2.1 PyBioS - Modeling and Simulation Platform

A modeling system for cellular reaction networks has to accomplish several requirements. It must have a well-defined internal structure for the representation of model components and reactions, and optionally functionalities for the storage of a model in a well defined structure, standardized format or database. Further desired aspects are a user-friendly interface for model development, a graphical representation of reaction networks, a detailed description of the mathematical model, integrated engines for deterministic or stochastic simulation along with graphical representations of their results, and functionalities for model analysis and model refinement. This is a very broad spectrum of functionalities.

Current modeling systems for biochemical research are usually designed for small- and medium-sized models. Most of them do not have functionalities for the visualization of the model's reaction network or are able to display only the entire topology, what makes the work with large models quite difficult. Furthermore, current modeling systems provide none or very rudimentary interfaces to major pathway databases, such as KEGG or Reactome. Latter point is extremely relevant, since the alternative to code computer models by hand is time-consuming and often error-prone.

To overcome these limitations I have invented the modeling and simulation system PyBioS that is described in the following in more detail.

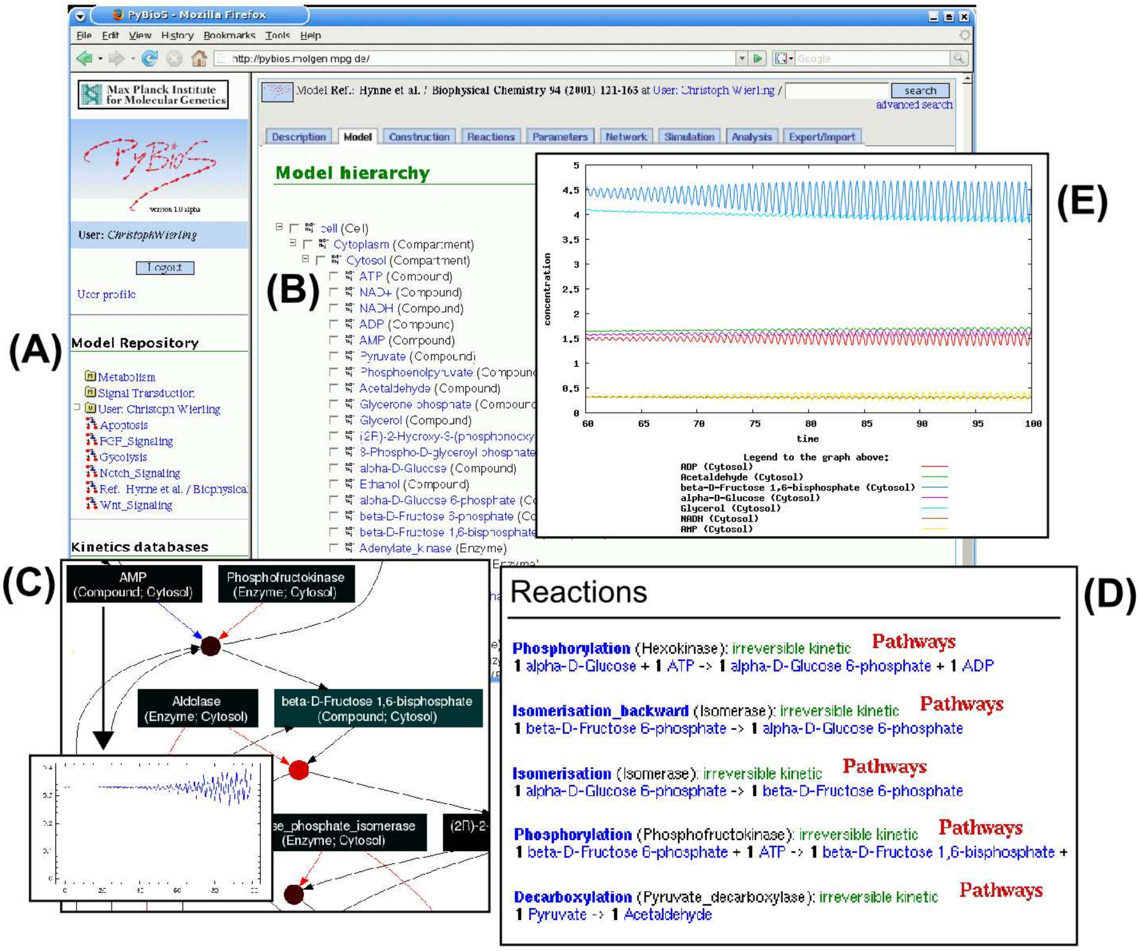
2.1.1 Overview of PyBioS

PyBioS is an object-oriented environment for the development and simulation of mathematical models of biological systems, which I have designed and developed at the Max Planck Intitute for Molecular Genetics (Wierling, 2006; Klipp et al., 2005; Wierling et al., 2007). It is designed as a software application for the World Wide Web[1]. Its user interface is depicted in Fig. 2.1. PyBioS serves as a hierarchical object oriented database to store models of cellular systems. Each model represents the model objects in a hierarchical object-oriented manner corresponding to cellular and molecular hierarchy. For instance, a model can hold a cell object that consists of a cytosol object and a nucleus object, where the cytosol compartment in turn can hold other objects, such as those representing proteins or other compounds like metabolites. Model objects are entities of the abstract *BioObject*-class that represents biological objects. Derived from this class are concrete classes for biological entities that are subdivided into container-like objects (*Environments*). The latter can contain other BioObjects and non container-like objects. This hierarchical structure is illustrated in Fig. 2.1A. Container-like object classes are Cell, Compartment, Complex and Chromosome. Non container-like object classes are Gene, pre-mRNA, mRNA, Polypeptide, Protein and Enzyme (of which also Polymerase, Spliceosome, RNase, Ribosome and Protease are derived from). Additional information such as annotation, sequence-data, parameters and initial concentrations are stored as object's properties. Actions, which describe reactions between different objects, are attached to BioObjects, e.g., a metabolic reaction is bound to its catalyzing enzyme.

Certainly, small subsystems can be modeled and analyzed to some extent in isolation by assuming steady and simplified boundary conditions. But as soon as these boundary conditions become variable—as given for a system as complex as the cell—it is clear that also this subsystem might behave differently in the context of a more comprehensive model. For instance ATP, one of the most important energy sources in the cell, is involved in diverse cellular processes and for example a massive consumption of ATP by a single process might have an important impact on other processes; this impact will not be discovered as long as the ATP concentration is handled as a constant parameter or as a variable of the isolated subsystem. Such constraints are, for example, also relevant for cellular signal transduction, where different signaling pathways can have an effect on each other through cross-talks.

Thus, PyBioS is particularly developed for the analysis of large models. Here, automated

[1] `http://pybios.molgen.mpg.de`


(A)
(B)
(C)
(D)
(E)
PyBioS - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://pybios.molgen.mpg.de/
Max Planck Institute for Molecular Genetics
PyBioS
version 1.0 alpha
User: ChristophWierling
Logout
User profile
Model Repository
Metabolism
Signal Transduction
User: Christoph Wierling
Apoptosis
FGF_Signaling
Glycolysis
Notch_Signaling
Ref. Hynne et al. / Biophysica
Wnt_Signaling
Kinetics databases
Model Ref.: Hynne et al. / Biophysical Chemistry 94 (2001) 121-163 at User: Christoph Wierling /
search
advanced search
Description
Model
Construction
Reactions
Parameters
Network
Simulation
Analysis
Export/Import
Model hierarchy
cell (Cell)
Cytoplasm (Compartment)
Cytosol (Compartment)
ATP (Compound)
NAD+ (Compound)
NADH (Compound)
ADP (Compound)
AMP (Compound)
Pyruvate (Compound)
Phosphoenolpyruvate (Compound
Acetaldehyde (Compound)
Glycerone phosphate (Compound
Glycerol (Compound)
(2R)-2-Hydroxy-3-(phosphonooxy
3-Phospho-D-glyceroyl phosphate
alpha-D-Glucose (Compound)
Ethanol (Compound)
alpha-D-Glucose 6-phosphate (C
beta-D-Fructose 6-phosphate (C
beta-D-Fructose 1,6-bisphosphate
Adenylate_kinase (Enzyme)
concentration
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
60
65
70
75
80
85
90
95
100
time
Legend to the graph above:
ADP (Cytosol)
Acetaldehyde (Cytosol)
beta-D-Fructose 1,6-bisphosphate (Cytosol)
alpha-D-Glucose (Cytosol)
Glycerol (Cytosol)
NADH (Cytosol)
AMP (Cytosol)
AMP (Compound; Cytosol)
Phosphofructokinase (Enzyme; Cytosol)
Aldolase (Enzyme; Cytosol)
beta-D-Fructose 1,6-bisphosphate (Compound; Cytosol)
se_phosphate_isomerase (Enzyme; Cytosol)
(2R)-2-

Reactions

Phosphorylation (Hexokinase): irreversible kinetic Pathways
1 alpha-D-Glucose + 1 ATP -> 1 alpha-D-Glucose 6-phosphate + 1 ADP

Isomerisation_backward (Isomerase): irreversible kinetic Pathways
1 beta-D-Fructose 6-phosphate -> 1 alpha-D-Glucose 6-phosphate

Isomerisation (Isomerase): irreversible kinetic Pathways
1 alpha-D-Glucose 6-phosphate -> 1 beta-D-Fructose 6-phosphate

Phosphorylation (Phosphofructokinase): irreversible kinetic Pathways
1 beta-D-Fructose 6-phosphate + 1 ATP -> 1 beta-D-Fructose 1,6-bisphosphate +

Decarboxylation (Pyruvate_decarboxylase): irreversible kinetic Pathways
1 Pyruvate -> 1 Acetaldehyde

Figure 2.1: PyBioS Web interface. The user can choose a particular model from the model repository **(A)** and inspect its hierarchical structure via the *Model*-tab **(B)** that also provides functionalities to edit the model. The *Network*-tab provides an automatically generated wiring diagram of the reaction network **(C)**, in which rectangular nodes represent the BioObjects (e.g., genes, compounds, proteins, etc.), and circular nodes the reactions (actions). The arrows in the diagram differ between mass-flow (black arrows) and information-flow (green and red arrows). The *Reactions*-tab lists all reactions of the model **(D)**. Via the *Simulations*-tab the user can run individual simulations. Time courses of the concentrations or fluxes are visualized graphically **(E)**.

import/export functions to populate models and an automatic generation of the mathematical model (ordinary differential equation system; ODE system) are essential.

PyBioS provides a broad spectrum of different functions for model design and development, simulation and analysis. It has different features that are outlined below and are introduced in the following sections.

- Object-oriented model design
- User interface for the development of individual models including interfaces to pathway databases such as KEGG, Reactome, and ConsensusPathDB
- Visualization of the model's network structure (model topology)
- Automatic generation of the deterministic mathematical model (ODE system)
- Numerical simulation using standard numerical integrators
- Methods for model analysis, like computation of conservation relations, detections of steady-states, stability analysis and parameter scan
- Repository of models and kinetics

2.1.2 Model structure

PyBioS employs an object-oriented strategy that was initially introduced by Stoffers et al. (1992). The authors used classes for metabolic entities and biochemical reactions for the modeling and simulation of metabolic systems.

Models in PyBioS have hierarchical object-oriented structures. Each model is stored in a separate *Model* object that contains the objects representing the biological entities. Biological entities, like genes, mRNAs, proteins, compounds, enzymes, complexes or compartments, are derived from the same class *BioObject*. BioObjects might have different properties. For instance, if a Michaelis-Menten kinetic is used, parameters like K_M or V_{max} are properties of the according *Enzyme* instance. Properties have a value, e.g., a floating point number. Properties might also be annotations, sequence-data, etc. Furthermore, one or several *Actions* can be bound to a BioObject. An action describes a biochemical reaction, physical process or a group of similar reactions or processes. Actions are described in more detail below. Fig. 2.2 gives an overview of the defined classes of biological objects and the information that is expected to be stored by these objects. Some of these BioObjects, like Compartment, Complex or Chromosome, are container-like objects (derived from the abstract class *Environment*)

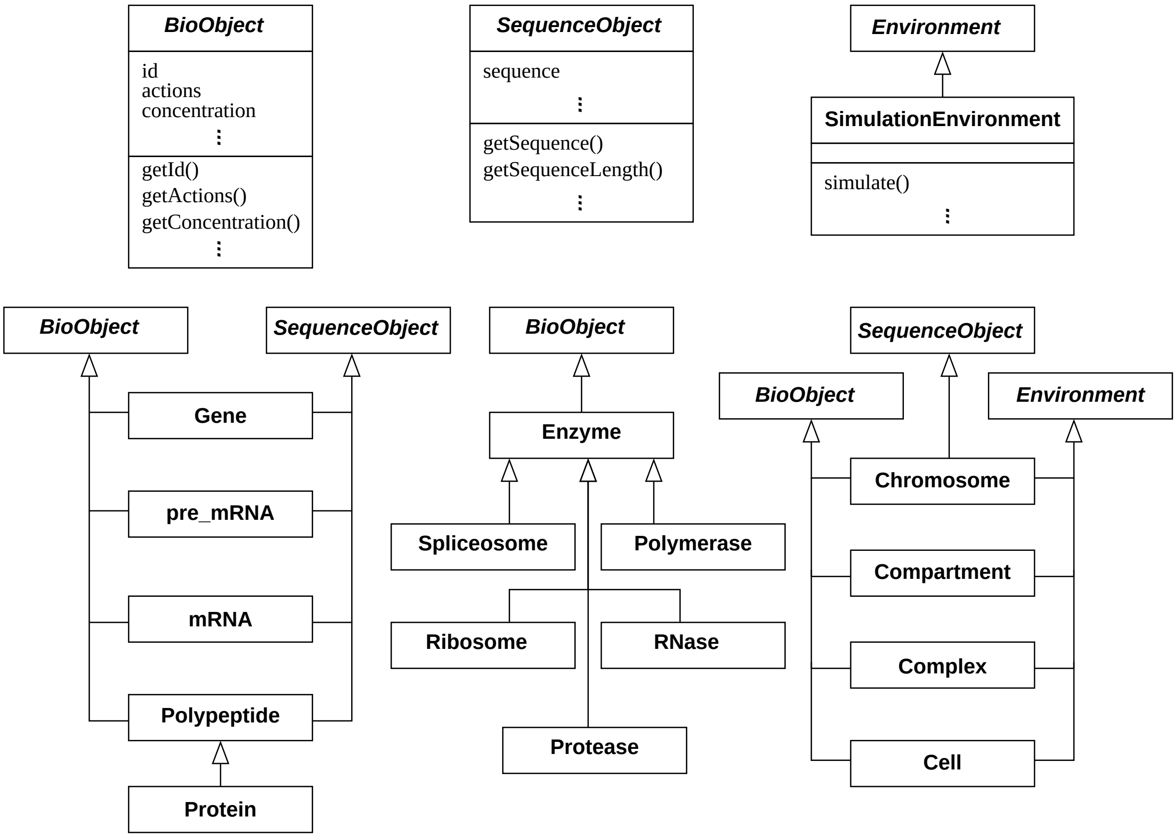


Figure 2.2: UML-diagram of the PyBioS ontology. According to UML notation, arrows point on classes from which other classes are derived. Central classes in the current version of PyBioS are the abstract classes *BioObject* and *Environment*. All classes which represent biological objects are derived from BioObject. BioObject has properties (in object-oriented programming denoted attributes) and methods (the diagram shows only some attribute and method examples). Properties are for instance the name of a BioObject (*id*) or its initial concentration (*concentration*). Methods are functions which belong to a certain class and operate on its attributes or other objects, that are passed along by the method call; e.g., *getId()* or *getActions()* return the object name or a list of the object's actions, respectively. Concrete classes, like *Gene*, *Enzyme* or *Cell* inherit from *BioObject* (and possibly other classes as well), which means that they have all the same attributes (but likely other values) and offer the methods of their parent class(es).

which can hold other BioObjects. All these classes define an ontology that is used for the internal representation of a model.

One or multiple actions can be assigned to a BioObject. Each action holds a directed reaction and a kinetic law (Fig. 2.3A). The directed reaction describes the mass flow from the substrate(s) to the product(s), as well as the molecularity (stoichiometry) with which they take part in the reaction.

Substrates and Products (denoted S and P, respectively) as well as the catalyzing enzyme (denoted E) are stored as lists associated to the action. The elements of these lists are references to BioObjects together with their respective stoichiometric coefficients. Further lists for modifiers (e.g., activators, inhibitors, etc.) can be added, if required. Reversible reactions are either constructed by defining the backward reaction as a separate action or by using a rate law that already considers this behavior and thus might become negative. In the latter case reversibility is indicated by a flag, which is an attribute of the used rate law. Kinetic laws are formulated in an abstract fashion by a list that consists of parameter and variable references and fundamental mathematical operations (+,-,*,/,log,exp, ...) and parentheses. The final kinetic law term is constructed from the lists of substrates, products, the enzyme, and other modifiers. The respective lists are used by predefined kinetic laws. A database of predefined kinetic laws is provided by PyBioS. Although no rule for assignment of actions to BioObjects is established, it makes sense for, e.g., an enzymatic reaction to attach it to the enzyme that catalyzes the given reaction. Autocatalytic reactions can be assigned to the substrate itself. Chemical reactions that take place in the absence of any catalyst can be assigned to the compartment-object they belong to or to a pseudo-object whose only task is to represent this action. An action that describes a transport process can be bound to a membrane- or transporter complex-instance.

From the individual rate laws of each action the ODE system is generated. This is described in Section 2.1.4.

2.1.3 Model Construction

As outlined in section 1.3, the development of a model of a biological system requires a lot of information, like information on the components of the system, the reactions they are involved in (topology and stoichiometry of the reaction network) and the kinetics of the individual reactions, which includes the kinetic laws and their respective kinetic parameters.

Comprehensive information about topology and stoichiomety of biological reaction systems is already available from suitable pathway databases. Unfortunately, information about reaction kinetics is limited, especially for large systems.

The first step of modeling is the collection of objects and reactions as well as appropriate

(A)

Action attributes	Representation of . . .	Example
Name (id):	Reaction identifier	e.g. Phosphorylation
Enzyme (E):	$E_1 \ldots E_k$	e.g. Hexokinase
Substrates (S):	$i_1\ S_1 \ldots i_l\ S_l$	ATP, Glucose
Products (P):	$j_1\ P_1 \ldots j_m\ P_m$	ADP, Glucose 6-phosphate
Modifiers (M):	$M_1 \ldots M_n$	
Reaction:	$i_1\ S_1 + \cdots + i_l\ S_l$ $\longrightarrow j_1\ P_1 + \cdots + j_m\ P_m$	ATP + Glucose $\longrightarrow$ ADP + Glucose 6-phosphate
Kinetics:	$v = f(S)$	$v = \frac{V_m[\mathrm{ATP}][\mathrm{Glc}]}{K_{\mathrm{DGlc}}K_{\mathrm{ATP}}+K_{\mathrm{Glc}}[\mathrm{ATP}]+K_{\mathrm{ATP}}[\mathrm{Glc}]+[\mathrm{Glc}][\mathrm{ATP}]}$

(B)

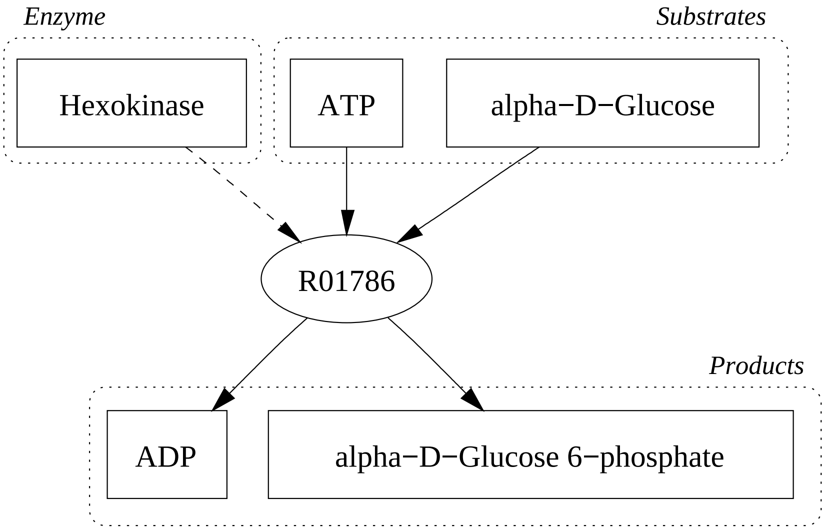


Figure 2.3: Description of an action. (A) Action data structure. E, S, P, and M are list of size k, l, m, n, respectively. These lists reference BioObjects that are involved in the reaction and their stoichiometric coefficients. **(B)** Graphical representation of the hexokinase action (dotted boxes show the elements of the substrate, product, and enzyme list, respectively).

kinetics that are relevant to the model. Using these data, a prototype can be developed. PyBioS provides three methods for model creation: (1.) via the Systems Biology Markup Language (SBML), (2.) via scripts using the application programmer interface (API) of PyBioS, or (3.) via the Web interface.

Import and export of models using SBML provides the ability to exchange models from PyBioS with other systems biology software applications. Moreover, this makes it possible to reuse existing SBML models as for example provided by the BioModels model repository (cf. Section 1.3.2).

The API that is implemented as a Web Service makes PyBioS very useful for other software applications, which do not include a simulation engine by themselves. For instance, PyBioS is used by the Gene Network Generator[2] (GeNGe) as an engine for the simulation of gene regulatory networks (Hendrik Hache, pers. comm.).

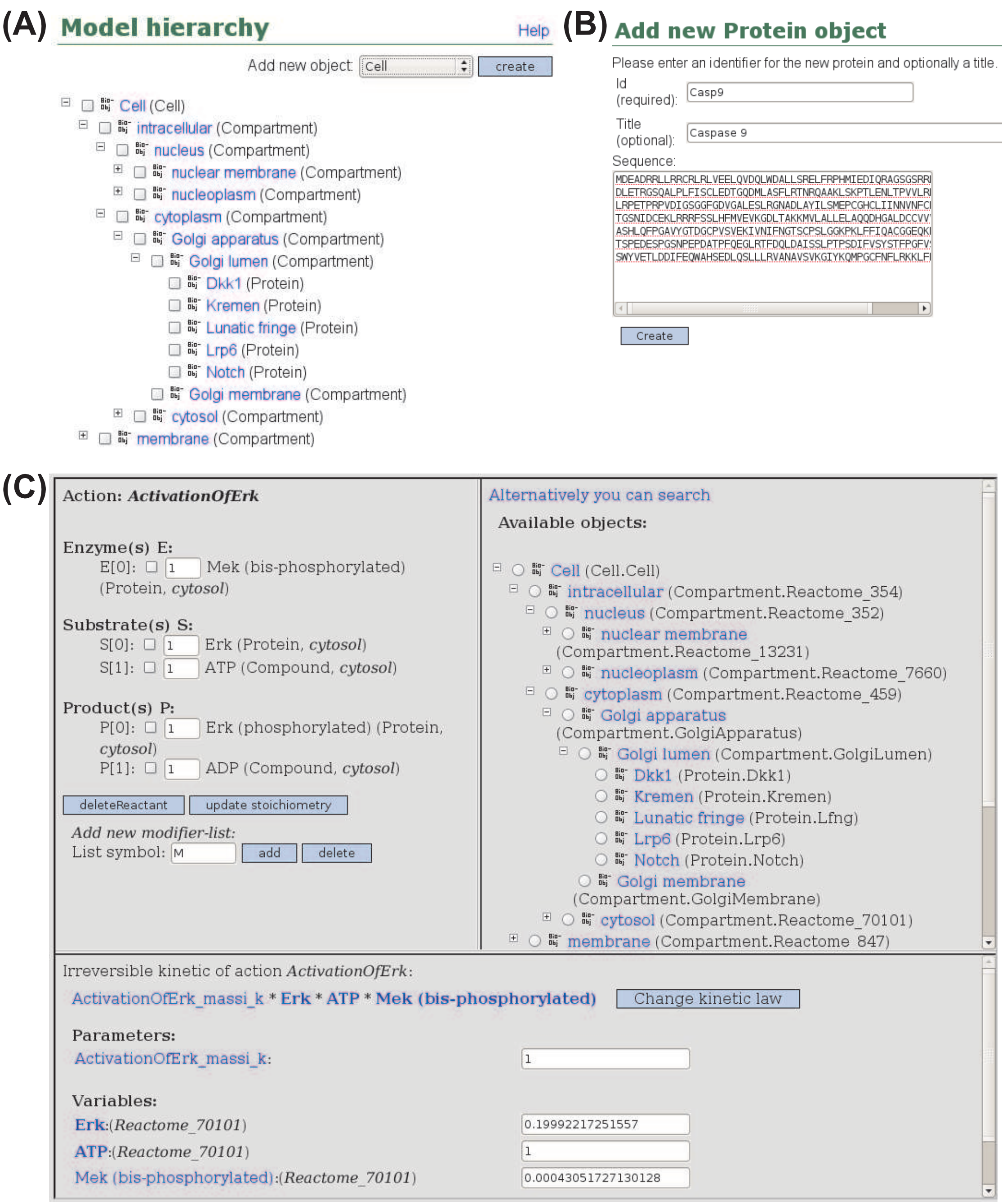
The third method to create a model in PyBioS is via its Web interface[3]. Model components and reactions can be added manually. (Fig. 2.4).

Alternatively to the development of an individual model by adding each reaction one by one, PyBioS has an interface to several major public pathway databases to retrieve pathway data automatically. The generic database interface of PyBioS accomplishes three major tasks: (1.) it provides a Web interface that enables the user to retrieve reaction-relevant data from the provided databases according to the users needs, (2.) it gathers the information of different databases in such a way that it can be used for the population of a single model, and (3.) it carries out the population of the PyBioS model by retrieving the relevant data from the according databases. The general database interface has references to the different database-specific low-level interfaces. Latter provide methods that are used by the general database interface for data retrieval, e.g., SQL-queries to the Reactome MySQL-database. Since BioObjects within PyBioS that were created using the database interface refer to their corresponding source database entry, additional information of the objects, such as diverse accession numbers, is still available. Low-level interfaces for the pathway databases Reactome, KEGG, and ConsensusPathDB are implemented (cf. Section 1.3.2).

The population of a model via the Web interface of the generic database adapter is simple. One can search for reactions of a specific gene or metabolite, or, alternatively, reaction of a specific pathway (Fig. 2.5A). From the list of results, the user can select several or all reactions to be created within the new model (Fig. 2.5B). The list of reactions that are selected for subsequent creation can be extended step by step with reactions of further database searches. Finally, the user can inspect the list of selected reactions again, choose appropriate kinetic laws from a predefined list (Fig. 2.6A), and instruct PyBioS to add those reactions to the cur-

[2] http://genge.molgen.mpg.de

[3] http://pybios.molgen.mpg.de



Model hierarchy

Add new object: Cell [create]

- ⊟ □ Cell (Cell)
  - ⊟ □ intracellular (Compartment)
    - ⊟ □ nucleus (Compartment)
      - ⊞ □ nuclear membrane (Compartment)
      - ⊞ □ nucleoplasm (Compartment)
    - ⊟ □ cytoplasm (Compartment)
      - ⊟ □ Golgi apparatus (Compartment)
        - ⊟ □ Golgi lumen (Compartment)
          - □ Dkk1 (Protein)
          - □ Kremen (Protein)
          - □ Lunatic fringe (Protein)
          - □ Lrp6 (Protein)
          - □ Notch (Protein)
        - □ Golgi membrane (Compartment)
      - ⊞ □ cytosol (Compartment)
  - ⊞ □ membrane (Compartment)

Add new Protein object

Please enter an identifier for the new protein and optionally a title.

Id (required): Casp9

Title (optional): Caspase 9

Sequence:

```
MDEADRRLLRRCRLRLVEELQVDQLWDALLSRELFRPHMIEDIQRAGSGSRR
DLETRGSQALPLFISCLEDTGQDMLASFLRTNRQAAKLSKPTLENLTPVVLR
LRPETPRPVDIGSGGFGDVGALESLRGNADLAYILSMEPCGHCLIINNVNFC
TGSNIDCEKLRRRFSSLHFMVEVKGDLTAKKMVLALLELAQQDHGALDCCVV
ASHLQFPGAVYGTDGCPVSVEKIVNIFNGTSCPSLGGKPKLFFIQACGGEQK
TSPEDESPGSNPEPDATPFQEGLRTFDQLDAISSLPTPSDIFVSYSTFPGFV
SWYVETLDDIFEQWAHSEDLQSLLLRVANAVSVKGIYKQMPGCFNFLRKKLF
```

[Create]

Action: *ActivationOfErk*

Enzyme(s) E:
E[0]: □ 1 Mek (bis-phosphorylated) (Protein, *cytosol*)

Substrate(s) S:
S[0]: □ 1 Erk (Protein, *cytosol*)
S[1]: □ 1 ATP (Compound, *cytosol*)

Product(s) P:
P[0]: □ 1 Erk (phosphorylated) (Protein, *cytosol*)
P[1]: □ 1 ADP (Compound, *cytosol*)

[deleteReactant] [update stoichiometry]

Add new modifier-list:
List symbol: M [add] [delete]

Alternatively you can search

Available objects:

- ⊟ ○ Cell (Cell.Cell)
  - ⊟ ○ intracellular (Compartment.Reactome_354)
    - ⊟ ○ nucleus (Compartment.Reactome_352)
      - ⊞ ○ nuclear membrane (Compartment.Reactome_13231)
      - ⊞ ○ nucleoplasm (Compartment.Reactome_7660)
    - ⊟ ○ cytoplasm (Compartment.Reactome_459)
      - ⊟ ○ Golgi apparatus (Compartment.GolgiApparatus)
        - ⊟ ○ Golgi lumen (Compartment.GolgiLumen)
          - ○ Dkk1 (Protein.Dkk1)
          - ○ Kremen (Protein.Kremen)
          - ○ Lunatic fringe (Protein.Lfng)
          - ○ Lrp6 (Protein.Lrp6)
          - ○ Notch (Protein.Notch)
        - ○ Golgi membrane (Compartment.GolgiMembrane)
      - ⊞ ○ cytosol (Compartment.Reactome_70101)
  - ⊞ ○ membrane (Compartment.Reactome_847)

Irreversible kinetic of action *ActivationOfErk*:

ActivationOfErk_massi_k * **Erk** * **ATP** * **Mek (bis-phosphorylated)** [Change kinetic law]

Parameters:

ActivationOfErk_massi_k: 1

Variables:

Erk:(*Reactome_70101*) 0.19992217251557

ATP:(*Reactome_70101*) 1

Mek (bis-phosphorylated):(*Reactome_70101*) 0.00043051727130128

Figure 2.4: Manual model generation in PyBioS. The hierarchical model structure **(A)** is established by adding individual BioObjects, e.g., a protein **(B)**, one by one. Reactions are added by assigning the participating objects to the respective lists, e.g., substrate or product list, and selection of an appropriate rate law from the kinetics repository **(C)**. Parameter values of the rate law and initial values of the participating objects can be assigned separately.

rent model (Fig. 2.6B). All objects and reactions populated by the generic database interface still provide references to their originating database entries. This feature enables an automated annotation of all model components in order to, e.g., associate the model components with their respective external association numbers. Additionally, the database references make it easier to merge models with other models and thus it supports the integration and re-usability of models.

In the following, the different features of the Web interface are outlined. Fig. 2.1 shows several features of the Web interface. A specific model can be inspected via different views that are accessible by several tabs. These views provide a representation of the hierarchical model structure, a listing of the model reactions, a graphical representation of the model network (a wiring diagram of the model), user interfaces for simulation and analysis and other functionalities. The "Construction tab" provides the access to the generic database interface and supports an easy model design. The user can search for reactions of a specific gene, metabolite or pathway (Fig. 2.5A). From the list of results the user can select several or all reactions for model population (Fig. 2.5B). The list of reactions that are selected for creation can be extended step by step with reactions of further database searches. All objects and reactions populated by the generic database interface still provide references to their originating database entries. This feature enables an automated annotation of all model components and makes it easy to extend the model by further database requests.

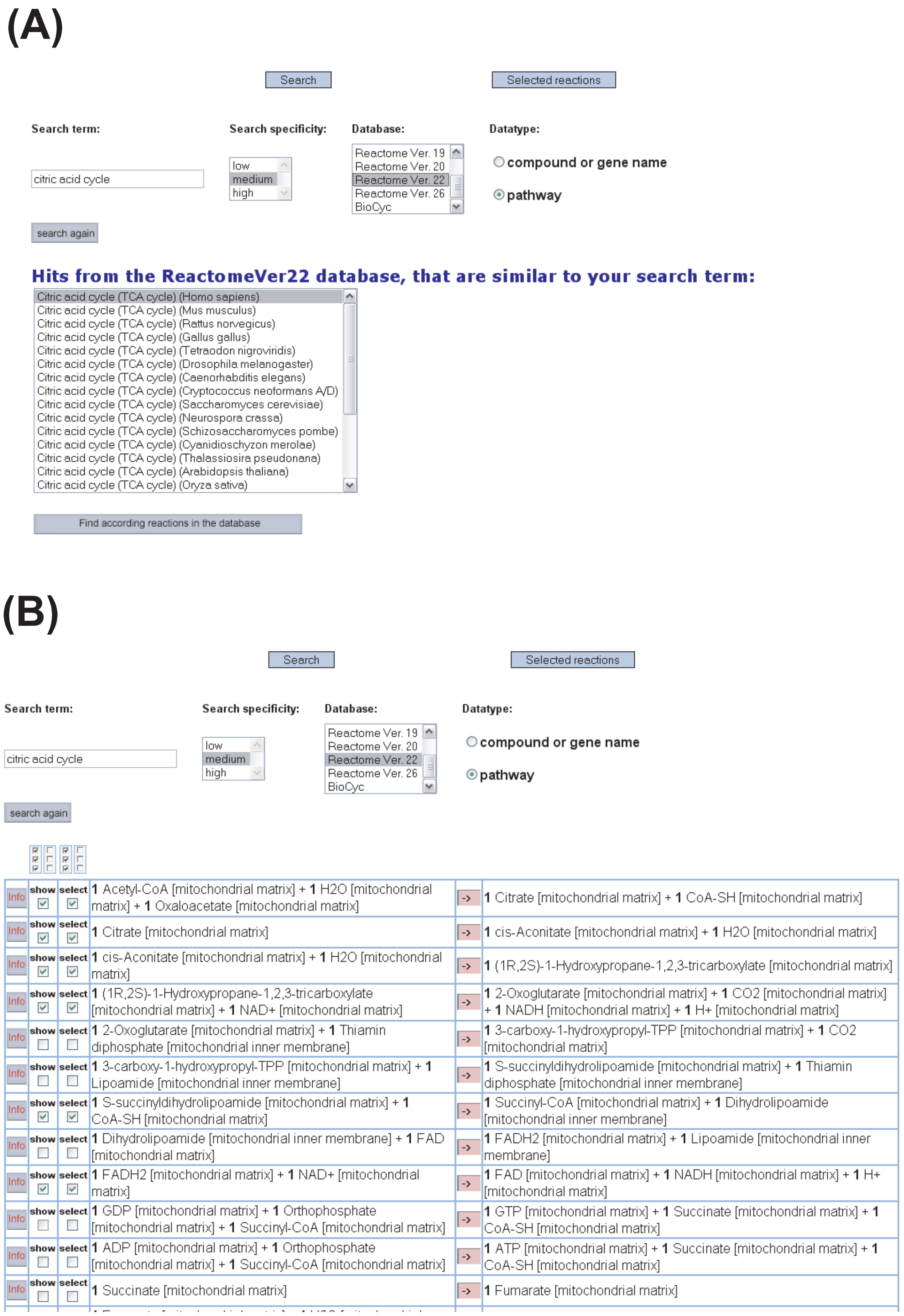
2.1.4 Quantitative Simulation

Reaction equations and rate laws that are defined by the actions, are used for the automatic generation of the ordinary differential equation (ODE) systems. The time change in the concentration of all species S_i is given by the following balance equation

$$\frac{\mathrm{d}[S_i]}{\mathrm{dt}} = \sum_{j=1}^{r} n_{ij} v_j(\mathbf{S}) \qquad i = 1, \ldots, m, \tag{2.1}$$

where r is the number of reactions, m the number of species, and n_{ij} the stoichiometric coefficient of S_i in the reaction j, which is positive for the production of S_i, negative for its degradation and otherwise defaults to zero. v_j denotes the velocity of reaction j, that is given by its rate law. $\mathbf{S}$ is a vector of concentrations of all species S_i.

PyBioS supports deterministic simulation by numerical integration of first order ODE-systems. It offers the use of the solvers LIMEX and LSODA (denoted SciPy in the modeling interface) to get the numerical solution of the initial value problem. LSODA (Hindmarsh, 1983; Petzold, 1983) is a solver for ordinary differential equations written in the programming language Fortran and a variant of the LSODE package. The algorithm used in this



(A)

Search | Selected reactions

Search term: citric acid cycle

Search specificity: low / medium / high

Database: Reactome Ver. 19 / Reactome Ver. 20 / Reactome Ver. 22 / Reactome Ver. 26 / BioCyc

Datatype: ○ compound or gene name ● pathway

search again

Hits from the ReactomeVer22 database, that are similar to your search term:

Citric acid cycle (TCA cycle) (Homo sapiens)
Citric acid cycle (TCA cycle) (Mus musculus)
Citric acid cycle (TCA cycle) (Rattus norvegicus)
Citric acid cycle (TCA cycle) (Gallus gallus)
Citric acid cycle (TCA cycle) (Tetraodon nigroviridis)
Citric acid cycle (TCA cycle) (Drosophila melanogaster)
Citric acid cycle (TCA cycle) (Caenorhabditis elegans)
Citric acid cycle (TCA cycle) (Cryptococcus neoformans A/D)
Citric acid cycle (TCA cycle) (Saccharomyces cerevisiae)
Citric acid cycle (TCA cycle) (Neurospora crassa)
Citric acid cycle (TCA cycle) (Schizosaccharomyces pombe)
Citric acid cycle (TCA cycle) (Cyanidioschyzon merolae)
Citric acid cycle (TCA cycle) (Thalassiosira pseudonana)
Citric acid cycle (TCA cycle) (Arabidopsis thaliana)
Citric acid cycle (TCA cycle) (Oryza sativa)

Find according reactions in the database

(B)

Search | Selected reactions

Search term: citric acid cycle

Search specificity: low / medium / high

Database: Reactome Ver. 19 / Reactome Ver. 20 / Reactome Ver. 22 / Reactome Ver. 26 / BioCyc

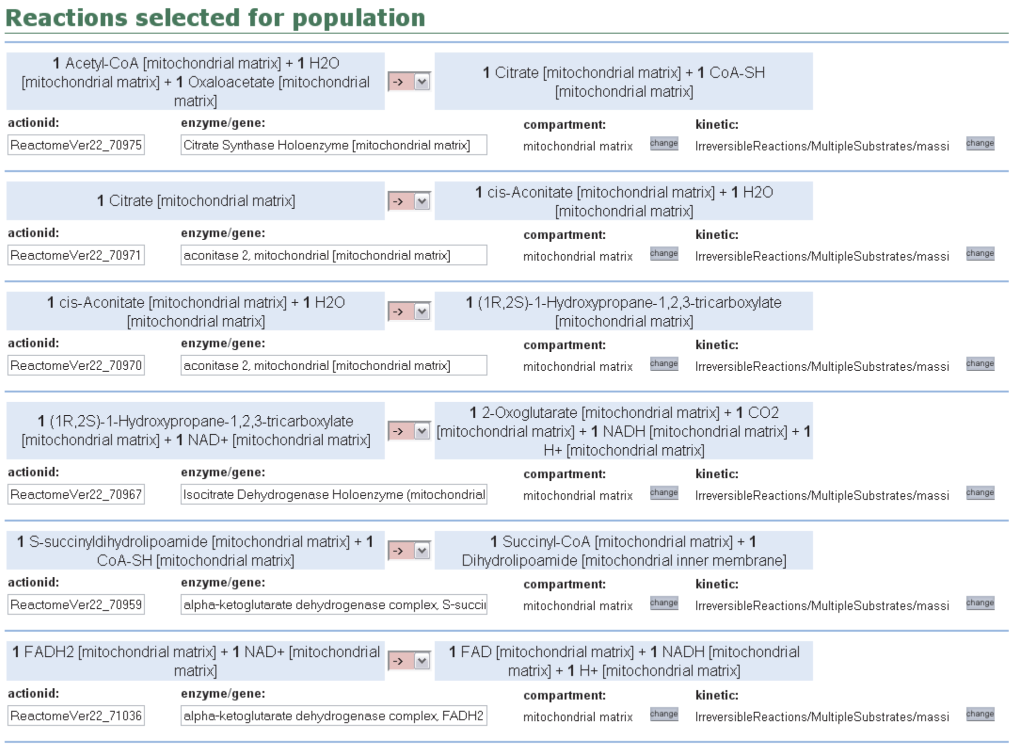
Datatype: ○ compound or gene name ● pathway

search again

Info	show ☑	select ☑	**1** Acetyl-CoA [mitochondrial matrix] + **1** H2O [mitochondrial matrix] + **1** Oxaloacetate [mitochondrial matrix]	->	**1** Citrate [mitochondrial matrix] + **1** CoA-SH [mitochondrial matrix]
Info	show ☑	select ☑	**1** Citrate [mitochondrial matrix]	->	**1** cis-Aconitate [mitochondrial matrix] + **1** H2O [mitochondrial matrix]
Info	show ☑	select ☑	**1** cis-Aconitate [mitochondrial matrix] + **1** H2O [mitochondrial matrix]	->	**1** (1R,2S)-1-Hydroxypropane-1,2,3-tricarboxylate [mitochondrial matrix]
Info	show ☑	select ☑	**1** (1R,2S)-1-Hydroxypropane-1,2,3-tricarboxylate [mitochondrial matrix] + **1** NAD+ [mitochondrial matrix]	->	**1** 2-Oxoglutarate [mitochondrial matrix] + **1** CO2 [mitochondrial matrix] + **1** NADH [mitochondrial matrix] + **1** H+ [mitochondrial matrix]
Info	show □	select □	**1** 2-Oxoglutarate [mitochondrial matrix] + **1** Thiamin diphosphate [mitochondrial inner membrane]	->	**1** 3-carboxy-1-hydroxypropyl-TPP [mitochondrial matrix] + **1** CO2 [mitochondrial matrix]
Info	show □	select □	**1** 3-carboxy-1-hydroxypropyl-TPP [mitochondrial matrix] + **1** Lipoamide [mitochondrial inner membrane]	->	**1** S-succinyldihydrolipoamide [mitochondrial matrix] + **1** Thiamin diphosphate [mitochondrial inner membrane]
Info	show ☑	select ☑	**1** S-succinyldihydrolipoamide [mitochondrial matrix] + **1** CoA-SH [mitochondrial matrix]	->	**1** Succinyl-CoA [mitochondrial matrix] + **1** Dihydrolipoamide [mitochondrial inner membrane]
Info	show □	select □	**1** Dihydrolipoamide [mitochondrial inner membrane] + **1** FAD [mitochondrial matrix]	->	**1** FADH2 [mitochondrial matrix] + **1** Lipoamide [mitochondrial inner membrane]
Info	show ☑	select ☑	**1** FADH2 [mitochondrial matrix] + **1** NAD+ [mitochondrial matrix]	->	**1** FAD [mitochondrial matrix] + **1** NADH [mitochondrial matrix] + **1** H+ [mitochondrial matrix]
Info	show □	select □	**1** GDP [mitochondrial matrix] + **1** Orthophosphate [mitochondrial matrix] + **1** Succinyl-CoA [mitochondrial matrix]	->	**1** GTP [mitochondrial matrix] + **1** Succinate [mitochondrial matrix] + **1** CoA-SH [mitochondrial matrix]
Info	show □	select □	**1** ADP [mitochondrial matrix] + **1** Orthophosphate [mitochondrial matrix] + **1** Succinyl-CoA [mitochondrial matrix]	->	**1** ATP [mitochondrial matrix] + **1** Succinate [mitochondrial matrix] + **1** CoA-SH [mitochondrial matrix]
Info	show □	select □	**1** Succinate [mitochondrial matrix]	->	**1** Fumarate [mitochondrial matrix]

Figure 2.5: Generic database interface (figure 1). (A) Search for citric acid cycle pathways (TCA). **(B)** Listing of the TCA related reactions available in Reactome.

(A)



Reactions selected for population

1 Acetyl-CoA [mitochondrial matrix] + **1** H2O [mitochondrial matrix] + **1** Oxaloacetate [mitochondrial matrix] -> **1** Citrate [mitochondrial matrix] + **1** CoA-SH [mitochondrial matrix]

actionid:	enzyme/gene:	compartment:		kinetic:	
ReactomeVer22_70975	Citrate Synthase Holoenzyme [mitochondrial matrix]	mitochondrial matrix	change	IrreversibleReactions/MultipleSubstrates/massi	change

1 Citrate [mitochondrial matrix] -> **1** cis-Aconitate [mitochondrial matrix] + **1** H2O [mitochondrial matrix]

actionid:	enzyme/gene:	compartment:		kinetic:	
ReactomeVer22_70971	aconitase 2, mitochondrial [mitochondrial matrix]	mitochondrial matrix	change	IrreversibleReactions/MultipleSubstrates/massi	change

1 cis-Aconitate [mitochondrial matrix] + **1** H2O [mitochondrial matrix] -> **1** (1R,2S)-1-Hydroxypropane-1,2,3-tricarboxylate [mitochondrial matrix]

actionid:	enzyme/gene:	compartment:		kinetic:	
ReactomeVer22_70970	aconitase 2, mitochondrial [mitochondrial matrix]	mitochondrial matrix	change	IrreversibleReactions/MultipleSubstrates/massi	change

1 (1R,2S)-1-Hydroxypropane-1,2,3-tricarboxylate [mitochondrial matrix] + **1** NAD+ [mitochondrial matrix] -> **1** 2-Oxoglutarate [mitochondrial matrix] + **1** CO2 [mitochondrial matrix] + **1** NADH [mitochondrial matrix] + **1** H+ [mitochondrial matrix]

actionid:	enzyme/gene:	compartment:		kinetic:	
ReactomeVer22_70967	Isocitrate Dehydrogenase Holoenzyme (mitochondrial	mitochondrial matrix	change	IrreversibleReactions/MultipleSubstrates/massi	change

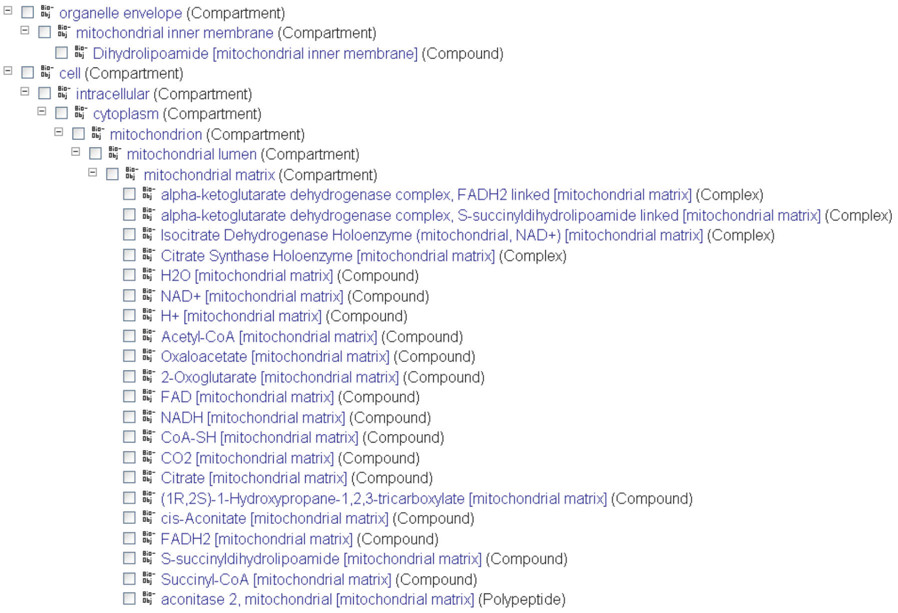
1 S-succinyldihydrolipoamide [mitochondrial matrix] + **1** CoA-SH [mitochondrial matrix] -> **1** Succinyl-CoA [mitochondrial matrix] + **1** Dihydrolipoamide [mitochondrial inner membrane]

actionid:	enzyme/gene:	compartment:		kinetic:	
ReactomeVer22_70959	alpha-ketoglutarate dehydrogenase complex, S-succi	mitochondrial matrix	change	IrreversibleReactions/MultipleSubstrates/massi	change

1 FADH2 [mitochondrial matrix] + **1** NAD+ [mitochondrial matrix] -> **1** FAD [mitochondrial matrix] + **1** NADH [mitochondrial matrix] + **1** H+ [mitochondrial matrix]

actionid:	enzyme/gene:	compartment:		kinetic:	
ReactomeVer22_71036	alpha-ketoglutarate dehydrogenase complex, FADH2	mitochondrial matrix	change	IrreversibleReactions/MultipleSubstrates/massi	change

(B)



- ⊟ □ organelle envelope (Compartment)
  - ⊟ □ mitochondrial inner membrane (Compartment)
    - □ Dihydrolipoamide [mitochondrial inner membrane] (Compound)
- ⊟ □ cell (Compartment)
  - ⊟ □ intracellular (Compartment)
    - ⊟ □ cytoplasm (Compartment)
      - ⊟ □ mitochondrion (Compartment)
        - ⊟ □ mitochondrial lumen (Compartment)
          - ⊟ □ mitochondrial matrix (Compartment)
            - □ alpha-ketoglutarate dehydrogenase complex, FADH2 linked [mitochondrial matrix] (Complex)
            - □ alpha-ketoglutarate dehydrogenase complex, S-succinyldihydrolipoamide linked [mitochondrial matrix] (Complex)
            - □ Isocitrate Dehydrogenase Holoenzyme (mitochondrial, NAD+) [mitochondrial matrix] (Complex)
            - □ Citrate Synthase Holoenzyme [mitochondrial matrix] (Complex)
            - □ H2O [mitochondrial matrix] (Compound)
            - □ NAD+ [mitochondrial matrix] (Compound)
            - □ H+ [mitochondrial matrix] (Compound)
            - □ Acetyl-CoA [mitochondrial matrix] (Compound)
            - □ Oxaloacetate [mitochondrial matrix] (Compound)
            - □ 2-Oxoglutarate [mitochondrial matrix] (Compound)
            - □ FAD [mitochondrial matrix] (Compound)
            - □ NADH [mitochondrial matrix] (Compound)
            - □ CoA-SH [mitochondrial matrix] (Compound)
            - □ CO2 [mitochondrial matrix] (Compound)
            - □ Citrate [mitochondrial matrix] (Compound)
            - □ (1R,2S)-1-Hydroxypropane-1,2,3-tricarboxylate [mitochondrial matrix] (Compound)
            - □ cis-Aconitate [mitochondrial matrix] (Compound)
            - □ FADH2 [mitochondrial matrix] (Compound)
            - □ S-succinyldihydrolipoamide [mitochondrial matrix] (Compound)
            - □ Succinyl-CoA [mitochondrial matrix] (Compound)
            - □ aconitase 2, mitochondrial [mitochondrial matrix] (Polypeptide)

Figure 2.6: Generic database interface (figure 2). **(A)** Listing of reactions selected for population. **(B)** PyBioS model automatically generated via the generic database interface.

solver switches between stiff and non-stiff methods automatically. PyBioS uses the interface to LSODA which is available from SciPy[4]. The solver LIMEX (Deuflhard et al., 1987; Deuflhard and Nowak, 1987) is an extrapolation integrator for the solution of linearly-implicit differential-algebraic systems (DAEs) written in Fortran[5]. It combines an implicit one step method with step size extrapolation to permit an adaptive control of step size and order.

PyBioS has an integrated interface for simulation ("Simulation tab"). One selects several or all components (reactions) and starts a simulation for a given time range. The time courses of the selected component concentrations (reaction fluxes) are subsequently plotted into a graph. Simulation results can also be plotted into a reaction network graph as described in the following paragraph.

2.1.5 Visualization

PyBioS model networks are defined by the BioObjects and their actions. Since a model of as few as 10 or 20 BioObjects becomes already very complex due to diverse substrates, products and modifiers, a visualization of the underlying model structure is of substantial benefit. Therefore, the biological network can be made more easily accessible by a graphical representation. An example of the visualization integrated in PyBioS is shown in Fig. 2.1C. In PyBioS, two kinds of nodes are used representing BioObjects (visualized by rectangles) or actions (visualized by circular nodes), respectively (Fig. 2.7). Relations between BioObjects and actions are visualized by directed arrows. The arrow color or style indicates either mass flow (black) or information flow (any other color or line style). The direction of the mass flow arrow indicates the mass flow, i.e. a BioObject is a substrate, if the arrow points from the BioObject node to the reaction node; otherwise, it is a product. Information flow arrows always pointing from BioObject nodes to action nodes, since they represent BioObjects, which catalyze or modify the particular action, but are consumed or produced by the respective reaction.

The PyBioS visualization interface can generate graphical representations of parts of the reaction network. Therefore, the user defines a subset of reactions to be displayed. This generates a network graph that can manually extended by clicking on one of the BioObject nodes and selecting further reactions of the selected BioObject that are not displayed in the current graph. In a similar way, reactions can also be removed from the current network graph. This feature provides a very flexible functionality for the inspection of the reaction network. In particular, this is very helpful for the inspection of large models. Furthermore, concentration values of the BioObjects or kinetic parameters of the reactions can also be

[4] http://www.scipy.org

[5] ftp://elib.zib.de/pub/elib/codelib/LIMEX4_2A1

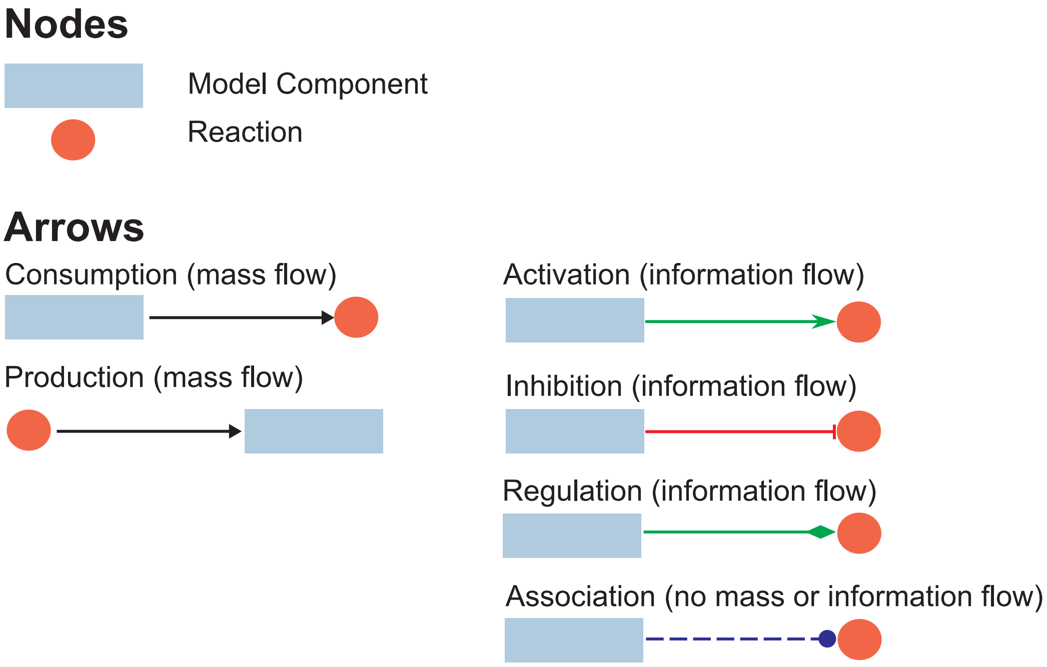

Nodes
Model Component
Reaction
Arrows
Consumption (mass flow)
Production (mass flow)
Activation (information flow)
Inhibition (information flow)
Regulation (information flow)
Association (no mass or information flow)

Figure 2.7: Elements for graphical representation in PyBioS. Node and arrow symbols used for the graphical representation of reaction networks within PyBioS.

modified within the graphical reaction network.

Moreover, simulation results can be displayed within the graph. For instance, reaction and BioObject nodes can be colored according to the simulation results of a specific time point. This highlights those BioObject that have very high or low concentration, or reaction that have a very fast or slow flux. Furthermore, the user can click on a specific node to display the time course simulation results of the component or flux.

2.1.6 Analysis Modules

The object-oriented model of the biological system as well as its derived mathematical model can be used for further analyses and consistency checks. For instance, it is possible to identify steady states, compute conservation relations and perform parameter scans.

Steady State Search

By definition, when the system has reached a steady state, the concentration of the metabolites does not change in time. The steady state of a system of reactions is characterized mathematically by $\mathrm{d}\mathbf{S}/\mathrm{d}t = \mathbf{0}$, where $\mathbf{S}$ is the vector of concentrations of the components and $\mathbf{0}$ is the null vector. In a nonlinear system this equation can have several solutions.

Two different methods for steady state search are available in PyBioS. First, an approach that depends on a root finding method to get the steady state using a numerical algorithm and second, a progressive simulation, that is called the *direct search*. Starting with simulation results of a user-defined time interval, the root finding approach computes the roots of the ODE system by using the MINPACK subroutine HYBRID1[6], which is a modification of the method described by Powell (1970).

The *direct search* performs a progressive time course calculation. Starting with the time interval $[t_0, t_n]$ specified by the user, a series $S_{t_n}, \cdots, S_{t_{n+k}}$ is calculated, where

$$S_{t_i} = \begin{pmatrix} S_{t_i,1} \\ \vdots \\ S_{t_i,m} \end{pmatrix} \qquad i = n, \cdots, n+k \tag{2.2}$$

is a vector with the concentrations of all species at time t_i. The *direct search* algorithm checks whether a steady state is reached by regarding

$$\|S_{t_{n+j+1}} - S_{t_{n+j}}\|_2 < \epsilon \qquad j = 0, \cdots, k-1 \tag{2.3}$$

for a user defined threshold ϵ. Here, $\| \cdot \|_2$ denotes the Euclidean norm. If this equation is satisfied for twenty consecutive time points in the interval $[t_n, t_{n+k}]$, it is assumed that the steady state is reached. Otherwise, another evaluation step starts for the interval $[t_{n+k}, t_{2n+k}]$. This is repeated up to l times (in the current version $l = 10$) until the steady state is found. In case of an unsuccessful search, the algorithm aborts and reports this.

Stoichiometric Matrix and Conservation Relations

Frequently, the amount of material of several molecular species involved in a cellular reaction network is conserved, e.g.,

$$\mathrm{n(ATP)} + \mathrm{n(ADP)} + \mathrm{n(AMP)} = \mathrm{const.} \tag{2.4}$$

where $n()$ denotes the amount of substance.

Such conservation relations can be computed by the network topology, which is given by the reactions and their stoichiometry. This topology of the reaction network describes the mass flow and is embodied in the stoichiometric matrix

$$\mathbf{N} = \begin{pmatrix} n_{11} & \cdots & n_{1r} \\ \vdots & \ddots & \vdots \\ n_{m1} & \cdots & n_{mr}. \end{pmatrix} \tag{2.5}$$

[6] http://www.netlib.org/minpack

A column of the matrix corresponds to a distinct reaction of the model and a row corresponds to a single molecular species (BioObject). r is the number of reactions and m is the number of species. An element $n_{ij} \neq 0$ indicates that a certain BioObject takes part in a particular reaction. The conservation matrix $\mathbf{\Gamma}$ can be obtained by computing the nullspace (kernel) of the stoichiometric matrix $\mathbf{N}$ using the relation $\mathbf{\Gamma N} = \mathbf{0}$.

Since it is conventional to compute the right nullspace, $\mathbf{\Gamma}^T$ (where T denotes the transposed matrix) is calculated from

$$\mathbf{N}^T \mathbf{\Gamma}^T = \mathbf{0} \tag{2.6}$$

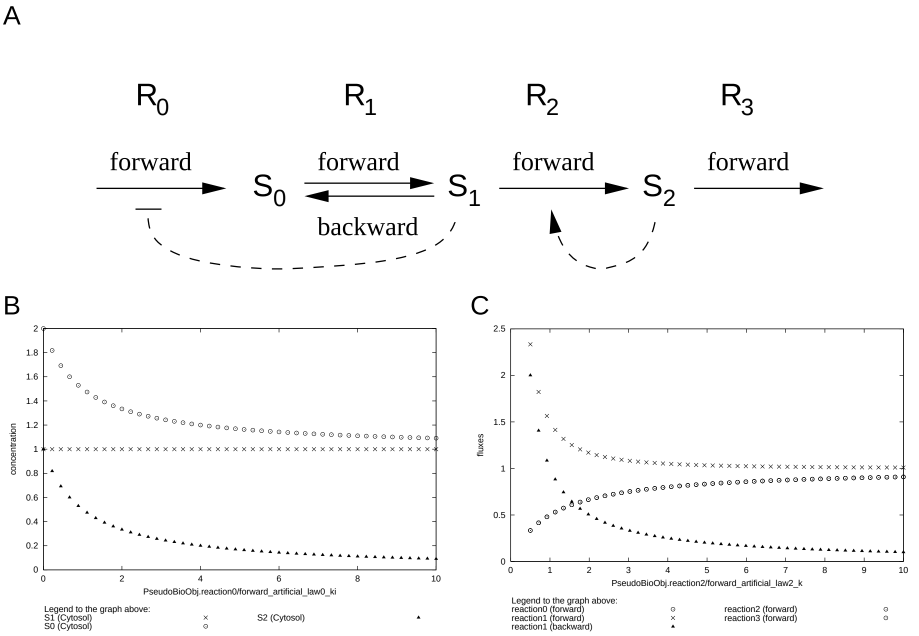
using the block diagonalization algorithm described by Schuster and Schuster (1991).

Parameter Scan

A parameter scan can be performed to analyze the behavior of the model. The possibility to consider the effect of one parameter on the concentrations of the metabolites and on the fluxes of the reactions is given by regarding the system in steady state. In steady state, the system is independent of time and an implicit dependence of the concentrations and fluxes on a parameter can be viewed. One parameter is varied in a given interval and the according steady states are computed by the direct search or root finding method. The results of the parameter scan (steady state concentrations or fluxes vs. the given parameter) are available as graphics or tab-delimited files. The parameter scan is illustrated by an artificial model shown in Fig. 2.8.

2.1.7 System’s Performance

Since molecular interaction data becomes massively available through the Internet and by rapidly evolving high-throughput techniques, strategies and methods for the integration of these data into biological models are required. Small systems of 20 or less objects can directly be translated into mathematical models by hand. However, the creation of models with several dozens, hundreds or even thousands of objects are not feasible anymore without an automation of this process. Therefore, the huge amount of experimental data as well as textbook data—which become increasingly available in a computationally amenable manner—are excellent sources for this purpose. PyBioS supports functionalities for the integration of external data sources. An interface to the metabolic data of the KEGG database enables the automated generation of models with a single or several pathways. Since the model creation is one central step in the process of model design, its scaling behavior is of interest. Therefore, metabolic models of different sizes in the number of reactions and objects were



Rate laws:

$$\begin{array}{rcl} R_0^{\rightarrow} & = & \dfrac{k_0}{1+k_i S_1} \\ R_1^{\rightarrow} & = & k_1 S_0 \\ R_1^{\leftarrow} & = & k_{-1} S_1 \\ R_2^{\rightarrow} & = & k_2 S_1 S_2 \\ R_3^{\rightarrow} & = & k_3 S_2 \end{array}$$

ODE-system:

$$\begin{array}{rcl} dS_0/dt & = & R_0^{\rightarrow} - R_1^{\rightarrow} + R_1^{\leftarrow} \\ dS_1/dt & = & R_1^{\rightarrow} - R_1^{\leftarrow} - R_2^{\rightarrow} \\ dS_2/dt & = & R_2^{\rightarrow} - R_3^{\rightarrow} \end{array}$$

Analytical solutions at steady state:

Concentrations:

$$\begin{array}{rcl} S_0 & = & \dfrac{1}{k_1}\left(\dfrac{k_0}{1+\frac{k_i k_3}{k_2}} + \dfrac{k_{-1}k_3}{k_2}\right) \\ S_1 & = & \dfrac{k_3}{k_2} \\ S_2 & = & \dfrac{1}{k_3}\left(\dfrac{k_0}{1+\frac{k_i k_3}{k_2}}\right) \end{array}$$

Fluxes:

$$\begin{array}{rcl} R_0^{\rightarrow} & = & \dfrac{k_0}{1+\frac{k_i k_3}{k_2}} \\ R_1^{\rightarrow} & = & \dfrac{k_0}{1+\frac{k_i k_3}{k_2}} + \dfrac{k_{-1}k_3}{k_2} \\ R_1^{\leftarrow} & = & \dfrac{k_{-1}k_3}{k_2} \\ R_2^{\rightarrow} & = & \dfrac{k_0}{1+\frac{k_i k_3}{k_2}} \\ R_3^{\rightarrow} & = & \dfrac{k_0}{1+\frac{k_i k_3}{k_2}} \end{array}$$

Figure 2.8: Parameter scan. (A) Reaction network of the artificial model. **(B)** A scan for parameter k_i in the interval [0,10] indicates that the concentration of S_1 is independent and S_0 and S_2 are dependent of this parameter. This is confirmed by its analytical solution. Similarly, flux changes can be analyzed **(C)**.

created using the interface to the KEGG database. This ranges from models with 20 objects and 11 reactions up to 1668 objects and 2365 reactions. For simplification, all reactions are considered to take place in the same compartment and are modeled by mass action rate laws. Kinetic parameters and initial concentrations are initialized with a value of 1. The CPU-time required for this model creation process was measured. Fig. 2.9A shows that the duration of the creation process scales linearly with the model size in this example with metabolic systems derived from the KEGG database. In parallel, the duration required for the simulation of the time-interval [0,10] and [0,1000] (arbitrary units) using the SciPy-solver was measured for each model. Here, a quadratic relation of time versus model size (given by the number of reactions) was found (Fig. 2.9B). It should be noted that the simulation time depends strongly on the complexity of the kinetic laws. The scaling behavior of some published models is also illustrated in Fig. 2.9B.

2.1.8 Summary of the Inventions

A modeling and simulation system for biochemical and cellular reaction networks called PyBioS was developed. PyBioS has a Web-based user interface for the creation of models and their subsequent simulation and analysis. Compared to other systems biology software applications, PyBioS has some unique features that make model development and simulation more efficient. It is, for instance, an interface to external pathway databases that makes it possible to import individual reactions, e.g., of a particular pathway into a PyBioS model. Moreover, PyBioS has a unique functionality to visualize results of a time course simulation along with the reaction network graph of the whole model or parts of it. Moreover, PyBioS provides several standard functions for the analysis of a model, such as computation of conservation relations, steady state search, or performing a parameter scan to evaluate the influence of a particular parameter on the steady state of the system. PyBioS was successfully used for the establishment of a molecular model of somitogenesis that is presented in the next section.

PyBioS has been selected as one of the top three contributions of the Heinz-Billing Award for Scientific Computation of the Max Planck Society in 2005.

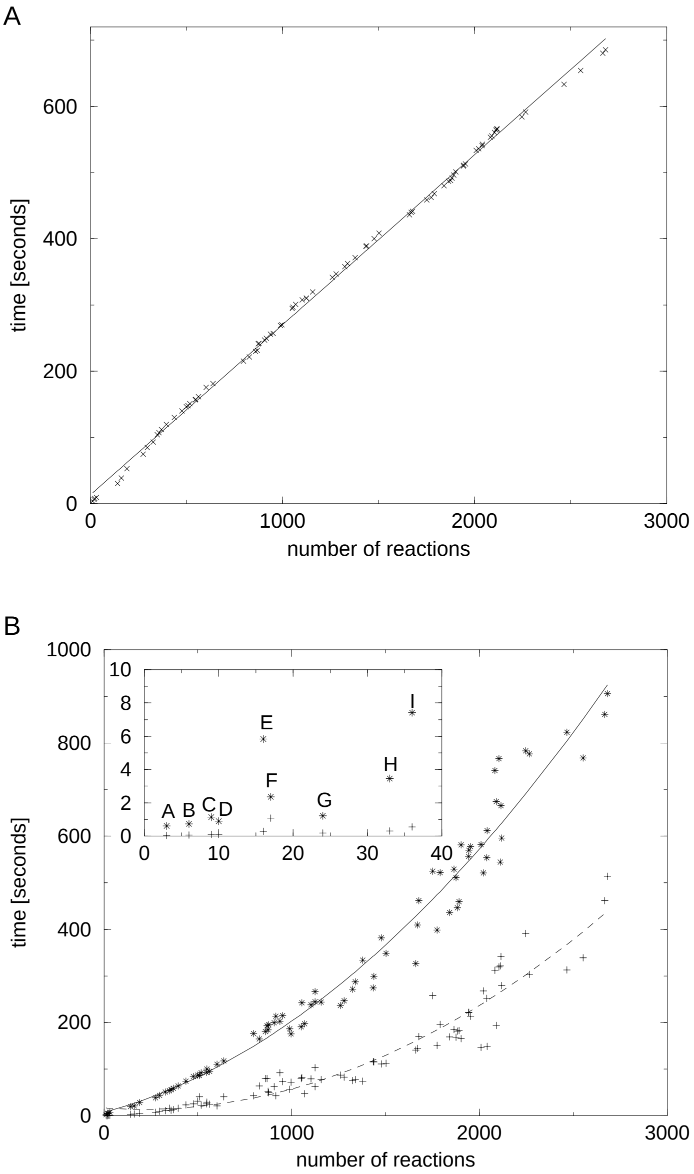


Figure 2.9: Scaling behavior of PyBioS for systems of different sizes. (A) Time required for the model creation; the straight line shows a linear regression. (B) Simulation for the time-interval [0,10] (+) and [0,1000] (*) using the numerical integrator of SciPy; the straight lines show quadratic regressions, respectively. The inserted graphic in B shows the scaling behavior of some models from the PyBioS models repository: (A) CellCycle-1991Tys-2, (B) CellCycle-1991Gol, (C) CellCycle-1991Tys, (D) MAPKcasc-2000Kho, (E) CircClock-2002Vil, (F) Metabolism-2000Teu, (G) CircClock-1999Lel, (H) CellCycle-1997Nov, (I) Hynne; (A)-(H) are imported via SBML from an SBML-model repository; (I) is described in Hynne et al. (2001).

2.2 Modeling of Biological Systems - Somitogenesis

As described in Section 1.2.1 somitogenesis is a general segmentation process taking place during vertebral embryogenesis. During somitogenesis epithelial blocks (the somites) regularly pinch off from the presomitic mesoderm (PSM) and eventually give rise to the axial skeleton, the skeletal muscles and the dermis of the back. It is assumed that the general regulatory mechanisms underlying somitogenesis are more or less similar across all vertebrates. There is evidence that somitogenesis is based on a molecular clock and a determination front established by a morphogenic gradient (cf. Section 1.2.1). For mouse, as well as several other mammals, it is proposed that the determination front is established by Wnt3a and Fgf8, two secreted signaling molecules that are produced in the tail bud and whose concentrations decay while the embryo elongates posteriorly (cf. Fig. 1.2; Aulehla and Herrmann 2004). The clock is assumed to be established by the signaling pathways Wnt, Notch, and FGF that are cross-linked with each other.

A general characteristic of somitogenesis is the regular formation of equally sized somites that sequentially pinch off from the PSM. The duration that it takes to form a single somite varies between different species, but is species-specific (cf. Tab. 1.1). Also species-specific is the number of vertebrae (vertebrae are derived from successive somites). It ranges from a few vertebrae in platyfish or frog to several hundreds in some cartilaginous fishes or long-bodied teleosts such as eels (Richardson et al., 1998).

On the molecular level several genes have been identified to oscillate during somitogenesis (cf. 1.2.1). This is, for example, in mouse and many other species, *Axin2* and *Dkk1*, components of the Wnt signaling pathway, *Lfng* and *Hes7*, which play a role in Notch signaling, and *Dusp6* and *Spry2*, which are known to be regulated by FGF signaling, but also by Notch signaling.

Different mathematical models describing the molecular processes underlying somitogenesis have already been proposed in the past (cf. Section 1.3.4). Several of them study the negative feedback regulation of the *Hes* gene by itself. But as outlined in Section 1.3.4 this autoregulatory mechanism is not sufficient for the description of molecular clock that controls somitogenesis. Goldbeter and Pourquié (2008) has developed a first model that integrates Notch, Wnt, and FGF signaling.

The mathematical model of somitogenesis that I have developed here is based on the conceptual model proposed by Aulehla and Herrmann (2004) and is adapted to current knowledge about the segmentation clock in mouse. It comprises several components known or assumed to be related with somitogenesis and being members of the Notch, Wnt and FGF signaling pathways. In the following I introduce two separate oscillatory models for the Notch and Wnt signaling pathways, respectively, and describe their individual features. Af-

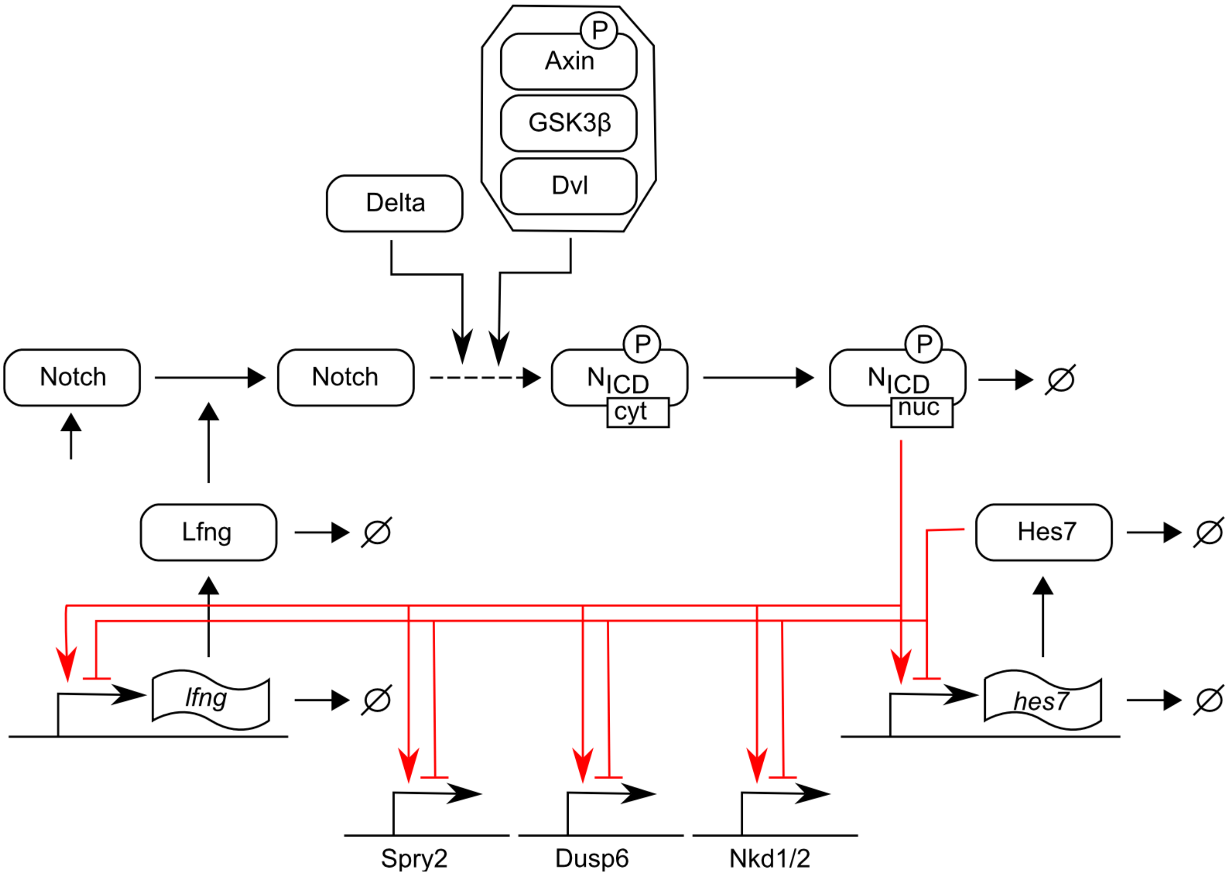


Figure 2.10: Notch model. Diagram of the Notch signaling pathway model.

terwards, cross-links between the individual signaling pathways and their connections to the FGF signaling are introduced and correlated with phenomenological aspects of somitogenesis. Kinetic parameters used within the models are extracted from Lee et al. (2003) and Goldbeter and Pourquié (2008) or, where no values were found in the literature, appropriate assumptions were used to reproduce the expected qualitative behavior.

2.2.1 Modeling Oscillatory Notch Signaling

The canonical Notch signaling pathway is described in Section 1.2.2.1. A general overview of the pathway is shown in Fig. 1.3. Key-players of the pathway are a Delta-type ligand and the Notch receptor. Once the Notch receptor is activated by the ligand, the Notch intracellular domain is cleaved off and can translocate into the nucleus and trigger the activation of target genes.

Using PyBioS a mathematical model of the Notch signaling pathway was implemented; its reaction network is depicted in Fig. 2.10. The model comprises the synthesis and post-translational modification of Notch, the release of the Notch intracellular domain (N_{ICD})

due to the activation via the Delta ligand and its phosphorylation by Axin:GSK3β:Dvl, its import into the nucleus, and the transcriptional activation of the target genes *Lfng*, *Spry2*, *Dusp6*, *Nkd1/2* and *Hes7* by nuclear $\mathrm{N_{ICD}}$. The Hes7 protein is a transcriptional inhibitor of itself as well as the other genes that are also under the control of $\mathrm{N_{ICD}}$ (*Lfng*, *Spry2*, *Dusp6*, *Nkd1/2*).

Necessary for oscillation of a molecular interaction network is a negative feedback loop (Tiana et al., 2007), whereas "negative feedback loop" simply defines a loop with an odd number of repressors. Besides this, a sufficient large time delay is also necessary to generate oscillations. This can be achieved, for example, by the introduction of a finite time delay, by a sharp response by some of the variables (e.g., described by a Hill kinetic as shown in Fig. 1.7A), or by a saturated degradation (e.g., described by a Michaelis Menten kinetic as shown in Fig. 1.6).

As described in Section 1.3.4 a molecular oscillator within the Notch signaling pathway can be established by *Hes7* whose protein is known to be a repressor of its own expression. It is described by the following ODE system.

$$\frac{d}{dt}[Hes7]^{\text{nucleoplasm}}_{\text{Pre\_mRNA}} = -v_0 + v_3 \tag{2.7}$$

$$\frac{d}{dt}[Hes7]^{\text{cytosol}}_{\text{mRNA}} = +v_0 - v_6 \tag{2.8}$$

$$\frac{d}{dt}[Hes7]^{\text{nucleoplasm}}_{\text{Protein}} = -v_1 + v_2 \tag{2.9}$$

$$\frac{d}{dt}[Hes7]^{\text{cytosol}}_{\text{Protein}} = +v_1 - v_2 - v_4 + v_5 \tag{2.10}$$

The rate laws of the different reactions are:

Hes7 mRNA export from nucleoplasm into cytosol

$$v_0 = k_0 \cdot [Hes7^{\text{nucleoplasm}}_{\text{Pre\_mRNA}}] \qquad \text{with} \quad k_0 = 0.1\text{min}^{-1} \tag{2.11}$$

Hes7 protein export form nucleoplasm into cytosol

$$v_1 = k_1 \cdot [Hes7^{\text{nucleoplasm}}_{\text{Protein}}] \qquad \text{with} \quad k_1 = 0.1\text{min}^{-1} \tag{2.12}$$

Hes7 import into the nucleoplasm

$$v_2 = k_2 \cdot [Hes7^{\text{cytosol}}_{\text{Protein}}] \qquad \text{with} \quad k_2 = 0.1\text{min}^{-1} \tag{2.13}$$

Transcription of Hes7

$$v_3 = V_3 \cdot (\frac{Ki_3{}^{ni_3}}{[Hes7^{\text{nucleoplasm}}_{\text{Protein}}]^{ni_3} + Ki_3{}^{ni_3}}) + a_3 \tag{2.14}$$

$$\text{with} \quad a_3 = 0.0\text{nM} \cdot \text{min}^{-1};\, V_3 = 0.2\text{min}^{-1};\, Ki_3 = 0.05\text{min}^{-1};\, ni_3 = 2.0 \tag{2.15}$$

Degradation of Hes7 protein

$$v_4 = Vm_4 \cdot \frac{[Hes7^{\text{cytosol}}_{\text{Protein}}]}{Km_4 + [Hes7^{\text{cytosol}}_{\text{Protein}}]} \tag{2.16}$$

$$Km_4 = 0.001;\, Vm_4 = 1.5 \tag{2.17}$$

Translation of Hes7 mRNA into protein

$$v_5 = k_5 \cdot [Hes7^{\text{cytosol}}_{\text{mRNA}}] \tag{2.18}$$

$$k_5 = 0.1\text{min}^{-1} \tag{2.19}$$

Degradation of Hes7 mRNA

$$v_6 = Vm_6 \cdot \frac{[Hes7^{\text{cytosol}}_{\text{mRNA}}]}{Km_6 + [Hes7^{\text{cytosol}}_{\text{mRNA}}]} \tag{2.20}$$

$$Vm_6 = 0.1;\, Km_6 = 0.01\text{min}^{-1} \tag{2.21}$$

The initial concentrations for an oscillatory state are:

$$[Hes7^{\text{nucleoplasm}}_{\text{Protein}}] = 0.473350\, nM$$
$$[Hes7^{\text{cytosol}}_{\text{mRNA}}] = 14.691751\, nM$$
$$[Hes7^{\text{cytosol}}_{\text{Protein}}] = 0.489142\, nM$$
$$[Hes7^{\text{nucleoplasm}}_{\text{Pre\_mRNA}}] = 0.215924\, nM$$

The implemented model is shown in Fig. 2.11. The model has an oscillatory behavior with a period of 110 min and shows, as expected, a time delay between the consecutive model components.

Another oscillatory circuit that can produce oscillations is the activation of the Notch receptor via Lfng, the subsequent expression of *Hes7* via N_{ICD}, and finally the negative feedback on *Lfng* expression by Hes7.

2.2.2 Modeling Oscillatory Wnt Signaling

The canonical Wnt signaling pathway is described in Section 1.2.2.2 (cf. also Fig. 1.4). A graphical illustration of the somitogenesis' Wnt signaling module is depicted in Fig. 2.12. A central component of the model is the destruction complex consisting of APC, Axin, and GSK3β that continuously phosphorylates β-catenin and thus targets it for degradation by the proteasome. The destruction complex is stabilized by the phosphorylated scaffold protein

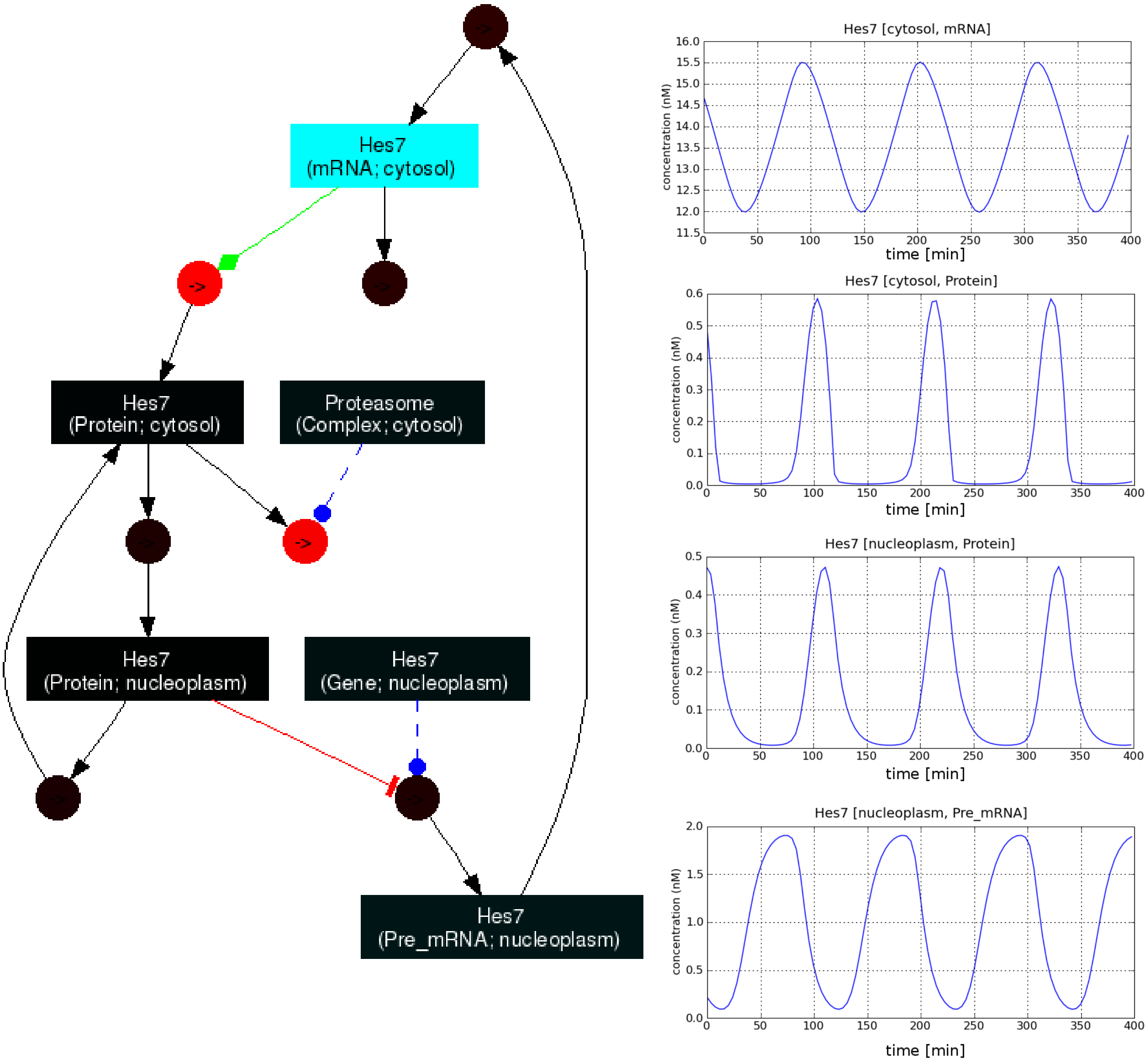


Figure 2.11: Model of Hes7 autoinhibition. (Left) Network of the reaction system; **(Right)** Simulation results of the oscillatory *Hes7* model.

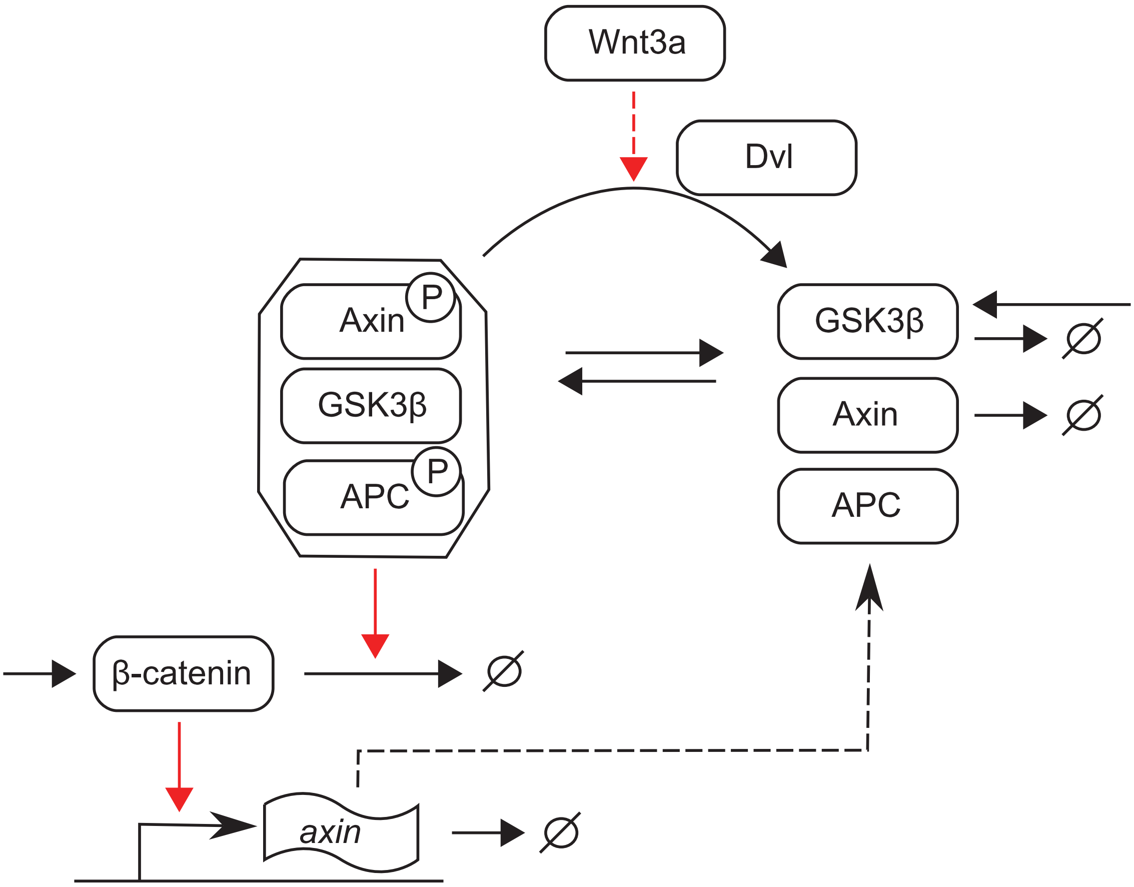


Figure 2.12: Wnt signaling module. Simplified illustration of the Wnt signaling model used within the somitogenesis model.

Axin whose expression is under the control of β-catenin. When Wnt signaling is activated by Wnt3a the destruction complex gets recruited to the plasma membrane by interaction with the activated Wnt/Frizzled receptor and Dvl. Subsequently, Axin is dephosphorylated and undergoes decay. This effect can be intensified by inhibition (phosphorylation) of the kinase GSK3β through activated Akt. β-catenin that acts as a co-activator for *Axin* expression is continuously synthesized. Without a Wnt3a signal its concentration is low, since it is continuously phosphorylated by the destruction complex and thus targeted for degradation. When the destruction complex is destabilized through a Wnt3a signal, the β-catenin concentration can increase and subsequently induce a delayed *Axin* expression that in turn results in the reformation of the destruction complex. Eventually, this leads to a decrease of the β-catenin concentration and of the *Axin* expression until the Axin concentration reaches a critical level and the cycle is restarted from the beginning. Unlike β-catenin and Axin, Dvl and GSK3β have a low turnover-rate so that β-catenin and Axin are the key players of this process.

To prove the described concept of a self-oscillating molecular clock established by the components of the Wnt signaling pathway, I have developed a corresponding mathematical model of this pathway with PyBioS. Fig. 2.13 shows a detailed graphical representation of the implemented model. Lee et al. (2003) developed a mathematical model of Wnt signaling in great detail, but without focusing on a potential oscillatory behavior of the pathway induced by a negative feedback as described above. For implementation that is presented here, I used several of the kinetic parameters from the model developed by Lee et al. Missing parameter values of the model were adapted to reproduce expected phenomenological findings.

Based on the developed model predictions for the description of Wnt signaling during somitogenesis were generated. It is known that in the posterior part of the PSM the Wnt3a concentration is high so that Wnt signaling can take place. Since Wnt3a is only produced in the tail bud of the embryo and undergoes a permanent decay, the Wnt3a concentration at a certain position within the PSM decreases continuously, while the embryo elongates at the tail. Once the Wnt3a concentration goes below a certain threshold value Wnt signaling arrests. Using the mathematical model, I have performed simulations for both system states, the "on" state and the "off" state of Wnt signaling. This was done by setting the external Wnt3a concentration to 1.0 nM and 0.0 nM, respectively. The obtained simulation results are shown in Fig. 2.14.

When Wnt signaling is activated by an external Wnt3a stimulus ("on" state of the Wnt signaling), a cyclic behavior of many components of the signaling pathway can be observed. This oscillatory behavior is a result of the delayed negative feedback loop that is established by the β-catenin controlled gene expression of Axin. The parameter set that was used for the simulation presented here generates an oscillation with a period of about 110 min. As the Wnt3a concentration declines (as it is observed in the PSM) the oscillation arrests (see Fig. 2.14 "Wnt signaling off"). For the "off" state, the concentration of the destruction complex (APC-P/Axin-P/GSK3β) is, compared to the "on" state, relatively high and, as a consequence of this, the β-catenin concentration is close to zero.

2.2.3 Coupling Wnt, Notch, and FGF signaling

A system of coupled oscillators underlying the segmentation clock has been proposed by Aulehla and Herrmann (2004); Dequéant et al. (2006) and Dequéant and Pourquié (2008). In contrast to a single autonomous oscillator, a system of coupled oscillatory networks might account for the robustness of the segmentation process.

Based on the models of Notch signaling and Wnt signaling an integrated model of both pathways was established. Furthermore, the integrated model was extended by components of FGF mediated signaling. Major components of the integrated FGF module are depicted in Fig. 2.15. It includes two pathways, one is the activation of Akt via the active FGF receptor

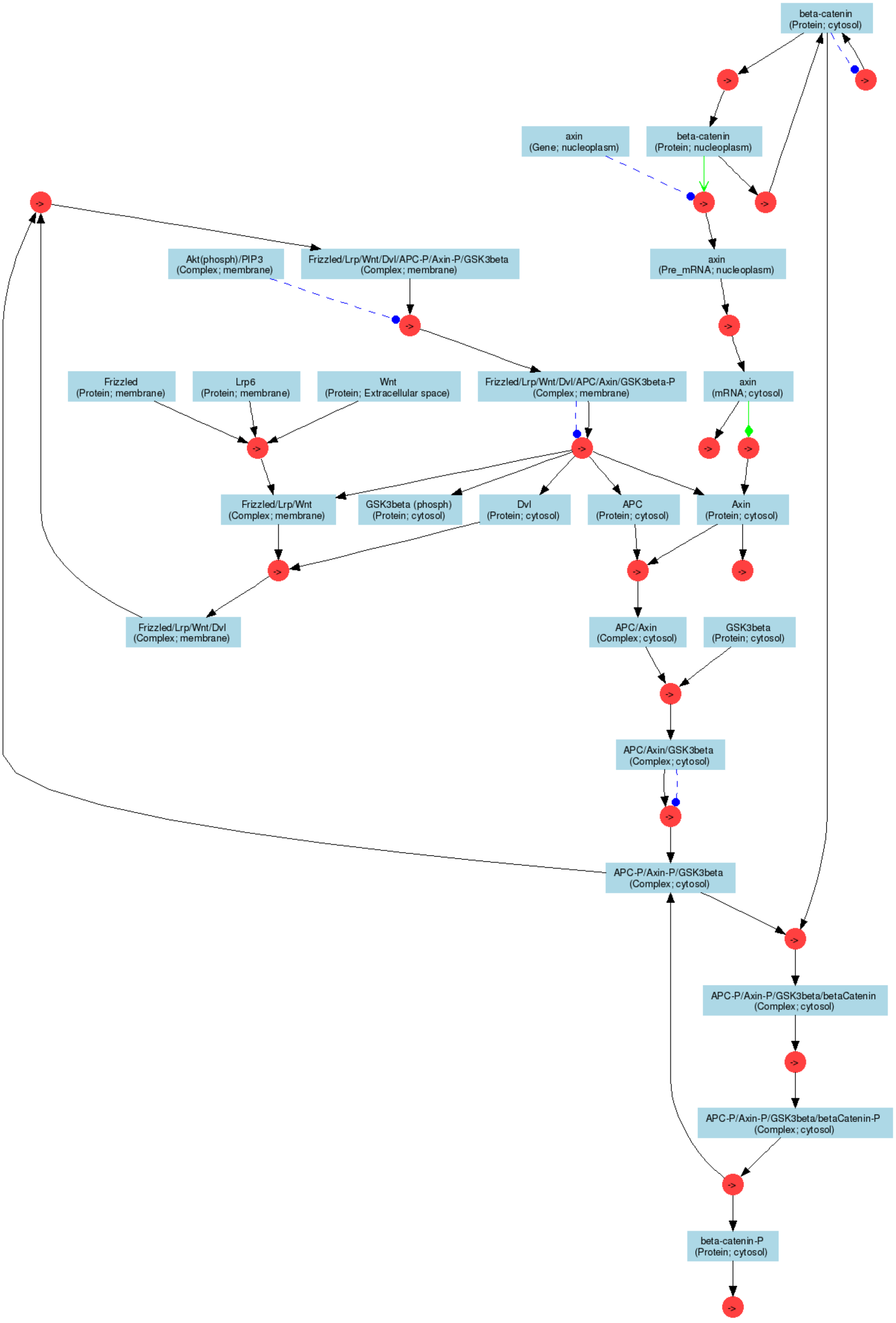

beta-catenin (Protein; cytosol)
axin (Gene; nucleoplasm)
beta-catenin (Protein; nucleoplasm)
Akt(phosph)/PIP3 (Complex; membrane)
Frizzled/Lrp/Wnt/Dvl/APC-P/Axin-P/GSK3beta (Complex; membrane)
axin (Pre_mRNA; nucleoplasm)
Frizzled (Protein; membrane)
Lrp6 (Protein; membrane)
Wnt (Protein; Extracellular space)
Frizzled/Lrp/Wnt/Dvl/APC/Axin/GSK3beta-P (Complex; membrane)
axin (mRNA; cytosol)
Frizzled/Lrp/Wnt (Complex; membrane)
GSK3beta (phosph) (Protein; cytosol)
Dvl (Protein; cytosol)
APC (Protein; cytosol)
Axin (Protein; cytosol)
Frizzled/Lrp/Wnt/Dvl (Complex; membrane)
APC/Axin (Complex; cytosol)
GSK3beta (Protein; cytosol)
APC/Axin/GSK3beta (Complex; cytosol)
APC-P/Axin-P/GSK3beta (Complex; cytosol)
APC-P/Axin-P/GSK3beta/betaCatenin (Complex; cytosol)
APC-P/Axin-P/GSK3beta/betaCatenin-P (Complex; cytosol)
beta-catenin-P (Protein; cytosol)
->

Figure 2.13: Wnt signaling model as implemented within PyBioS.

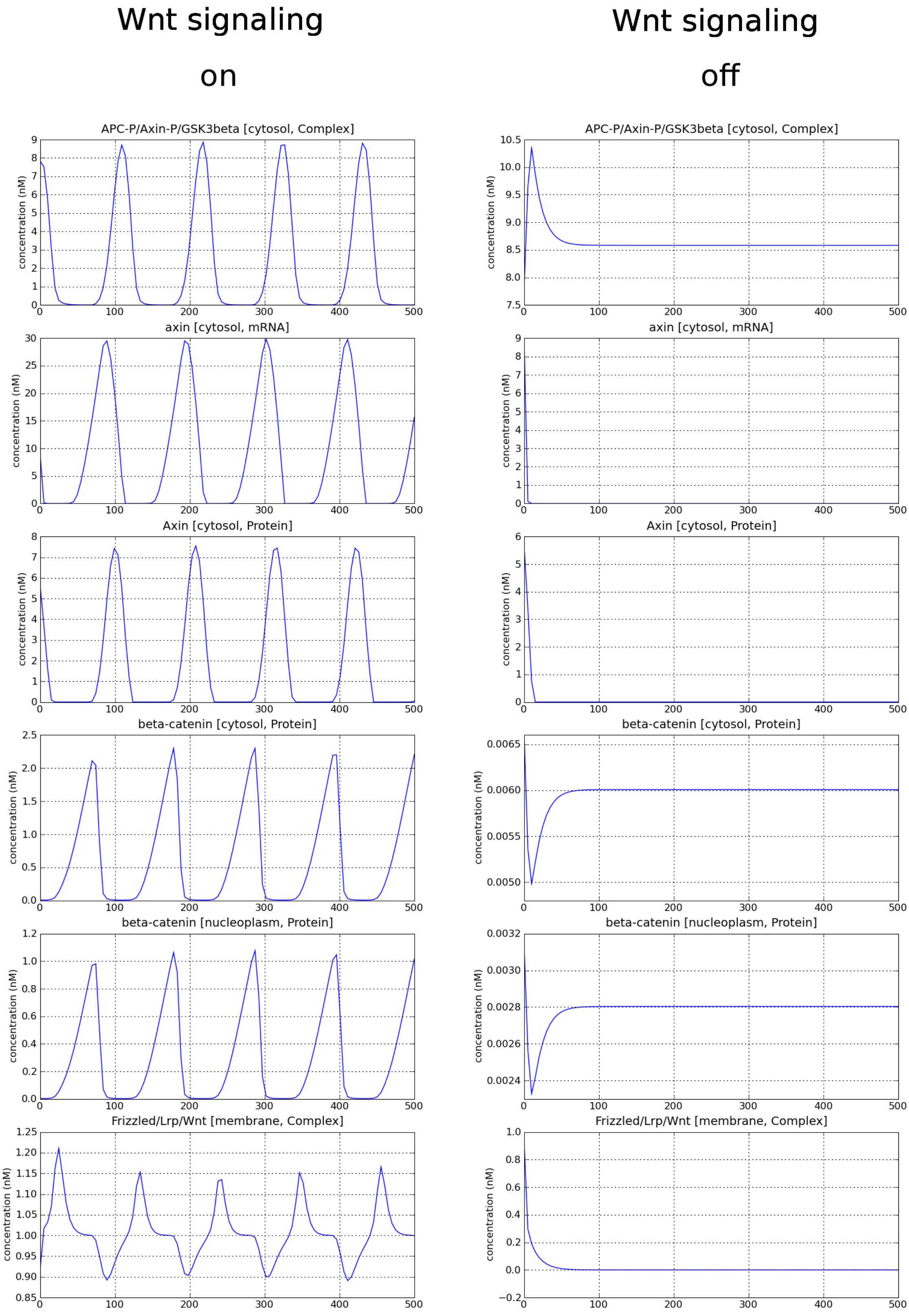


Figure 2.14: Simulation results of Wnt signaling. When Wnt signaling is on (Wnt3a is present) oscillations can take place. Without an extrinsic Wnt3a signal the β-catenin concentration is low and the oscillation stops.

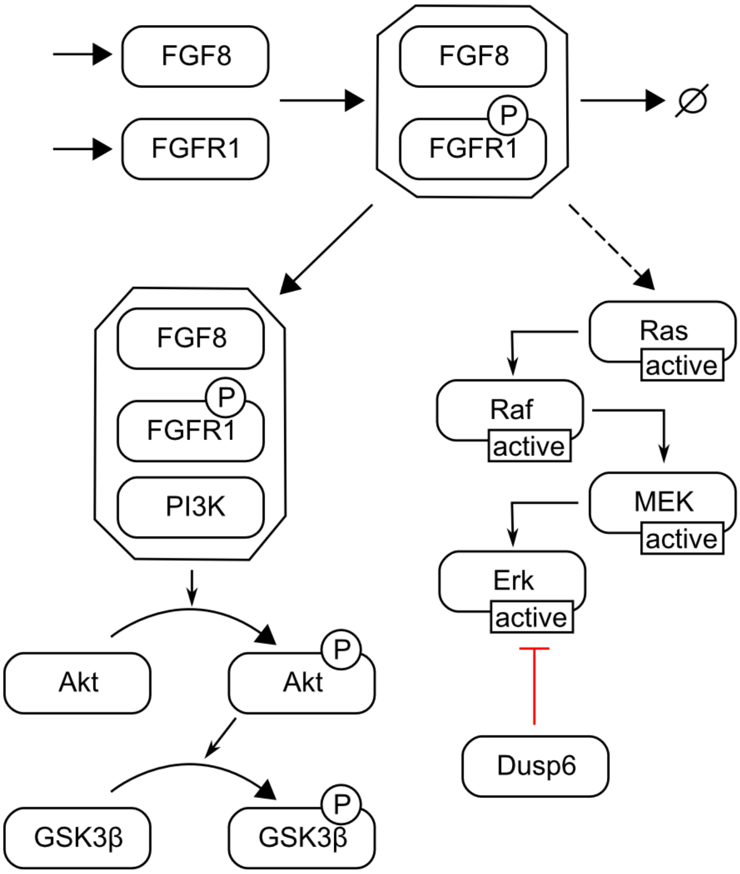


Figure 2.15: FGF model. Diagram of the FGF signaling pathway elements.

and PI3-kinase. The other is the MAPK pathway via the FGF receptor (SOS, Grb2, and Frs2 are also include, but not shown in Fig. 2.15), Ras, Raf, Mek, and finally Erk.

Different cross talks between the Notch, Wnt, and FGF signaling are also implemented in the integrated model. One cross talk is establish between active Akt and GSK3β. In this interaction GSK3β can become phosphorylated and by this inhibited. A second link is the positive regulation of *Dusp6* expression by active Erk. However, Dusp6 is a phosphatase that can dephosphorylate Erk. This negative feedback of Dusp6 on Erk can also account for an oscillator.

The integrate model consists of 118 components and 161 reactions.

2.3 Modeling of Laboratory Methods - DNA Array Experiments

Besides the analysis of biological systems, modeling and simulation strategies can also be applied to biotechnological experimental techniques. One particular technique of high interest in molecular genetics is gene expression analysis using DNA array technology. An introduction to DNA array technology is given in Section 1.4.1.

Summarizing, DNA array technology is based on the hybridization of labeled ssDNA to its complementary strand called probe. Different probes are fixed as spots on planar surfaces, like glass slides or nylon filters. The experimental data that was used for the presented model originates from cDNA array experiments spotted on nylon filters, but the presented approach can also be applied to arrays based on glass slides, since the problems for the quantification and statistical evaluation are very similar. Crucial for DNA experiments is the reliability of the produced data and their reproducibility. To ensure both reliability and reproducibility a sophisticated experimental design is necessary. This includes the identification of error parameters that affect the hybridization data during the data generation process. Influences of systematic and statistical errors due to biotechnological methods (for example mRNA preparation, PCR, hybridization) as well as due to devices and array media (for example robots, filters, glass slides) and their effects on evaluation software and algorithms (image analysis, statistical tests) must be estimated. I have developed a computer simulation that takes into account several sources of error, such as variations of spot shapes, spot positions, and local and global background noise. The simulation environment was used to judge the influence of these parameters on subsequent data analysis, for instance image analysis and the detection of differentially expressed genes. The presented model and simulation study was published in BMC Bioinformatics (Wierling et al., 2002).

The hybridization signal intensities that were used as input data for the simulation study is taken from experimental data. The data was derived as mean values from six cDNA nylon filters each of which was spotted with the same set of 14208 zebrafish cDNA clones and each was hybridized independently with the same complex target of an mRNA pool obtained from zebrafish gastrula stage embryos. The output are series of filter images containing well-defined error parameters. In each series only a single parameter was varied at once in order to measure its effects on data analysis. The range of parameter variation was adapted to real experiments that were used as experimental reference for the simulations.

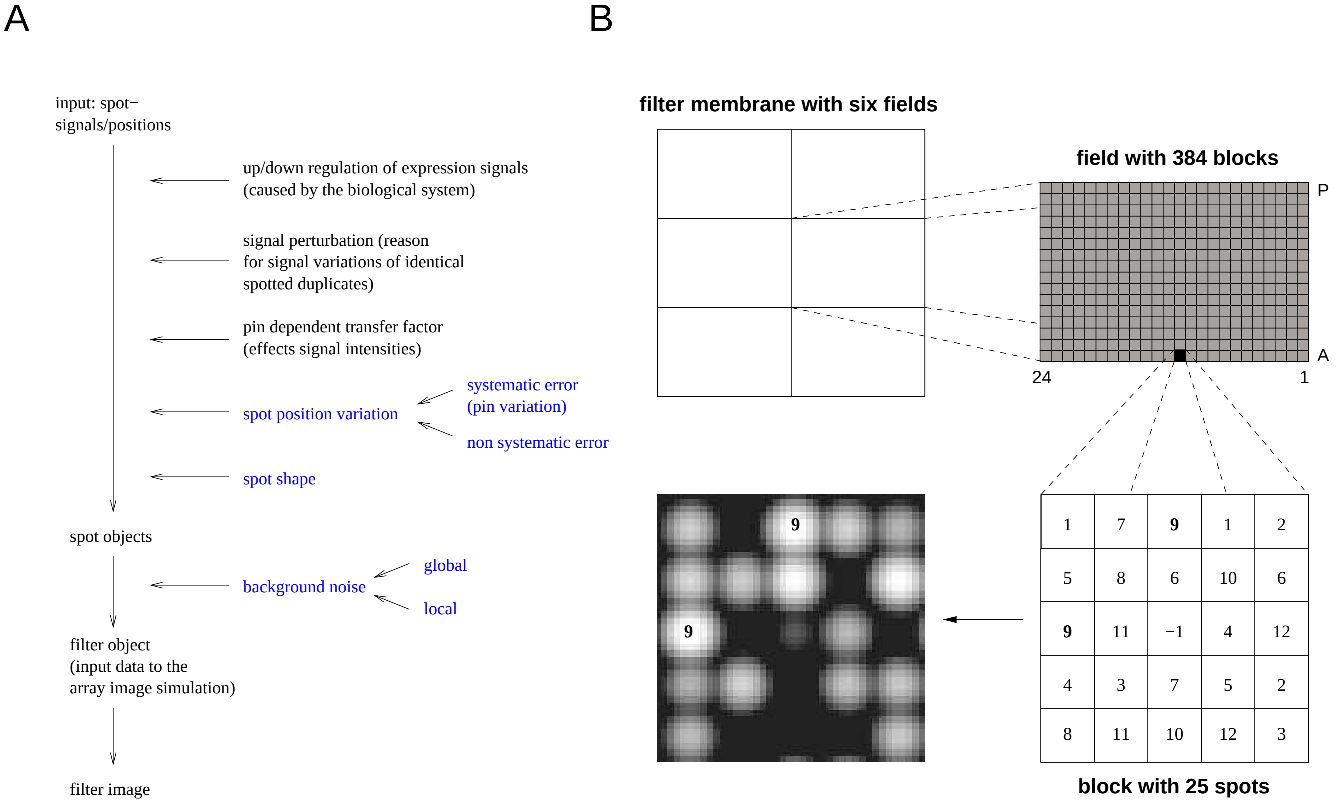
After creating the simulated data, the effect of the error parameters on the subsequent data analysis pipeline was measured. Two modules of this pipeline are highlighted: Image analysis and statistical analysis of differentially expressed genes, although the simulation tool is not restricted to these applications. I chose image analysis because it is the first module of the data analysis and builds the basis for all further research and statistical analysis of differentially expressed genes because it is one of the most utilized applications of gene arrays.

The images were analyzed with three different image processing programs. Parameters that are judged in this study are variations of the spot positions caused by different experimental artifacts and different sources of background noise. For gene expression profiling twelve filters with varying local background and experimentally determined signal variations were simulated, six of them correspond to hybridizations with a *treatment* and six of them correspond to hybridizations with a complex *control* target. I analyzed how many experimental repetitions are necessary to detect a given level of differential expression. The significance of the differential expression was judged by P values computed by the Welch t-test (cf. Herwig et al., 2001).

2.3.1 Implementation of the Simulation Tool

The simulation tool is written in the object-oriented scripting language Python. Some computation intensive functions are implemented in the programming language C and can be used as modules in Python. Objects like filters, spots or hybridization-data are stored as persistent objects by the use of Zope[7]. Fig. 2.16A illustrates the implemented simulation pipeline. It takes a set of expression data as input (I used an experimental signal distribution of hybridization data, see section A.3.1) and their position on the array. During the simulation pipeline several perturbations can be performed. Signal intensities can change due to the up- or down-regulation of gene expression, independent perturbations that effect signal differences of identically spotted duplicates can arise, or a systematic error happens during the spotting process due to pin-dependent differences in the amount of transfered PCR-product. Perturbations of systematic or non-systematic spot position errors and varying spot shapes are also considered. These perturbations result in the input data (filter object, which references its spot objects) used for the array image simulation. Depending on the type of array (filter or glass slide) different levels of global or local background noise can be considered here. The simulation parameters that are under investigation in this study are listed in Tab. 2.1. The output of one array simulation is a parameter file (that contains the values of the variation parameters), a file with the input data for the array image simulation (that contains signal and background intensities and the spot positions) and the image itself as a 16 bit Tiff-file.

[7] `http://www.zope.org/`



1	7	**9**	1	2
5	8	6	10	6
9	11	−1	4	12
4	3	7	5	2
8	11	10	12	3

Figure 2.16: Simulation pipeline and array layout. (**A**) Diagram of the filter simulation pipeline. The parameters highlighted in blue are the parameters that were varied (cf. Tab. 2.1). (**B**) Layout of a filter membrane with 57 600 spot positions. A 5×5 spotting pattern is shown; spots with identical position numbers (e.g. No. 9) indicate duplicates. -1 denotes a constant anchor spot which is identical for each block.

2.3.2 Data Sets

The quality of an expression analysis strongly depends on the distribution of the signal intensities and the spot positions on the filter (e.g., outshining effects). To deal with a realistic situation, results of real experiments were used as input data for the construction of the artificial data and the statistical expression analysis.

2.3.2.1 Design of Artificial Sample Sets

In order to detect differentially expressed genes with an experimental setup, the cDNA clone array is hybridized with two mRNA targets of different origin: one target commonly originates from a reference tissue (’control’), the second target originates from treated tissue, where ’treated’ refers to a certain chemical treatment, a mutant or a disease (’treatment’).

In this simulation setup the signals for the control target hybridization were taken from a signal-distribution derived from corresponding experimental data of 14 208 clones (see

Table 2.1: Definition, modeling, and critical effects of simulation parameters.

Parameter	Model	Variation	Critical effect[1]
Spot variation	spot shift (Gaussian distribution)	SD from ideal position	SD $>$ 0.15–0.2 mm $\hat{=}$ 16.7–22.2 %[2]
Pin variation	block shift (Gaussian distribution)	SD from ideal position	SD $>$ 0.12–0.167 mm $\hat{=}$ 13.3–18.6 %[2,3]
Spot shape	a) two-dimensional Gaussian distribution b) Crater spot distribution c) Plateau spot distribution	a) no variation (fixed SD $= 0.1482$ mm) b) radius of crater c) no variation (fixed radius of cylindric plateau spot $= 0.342$ mm)	b) radius $>$ 0.1995 mm $\hat{=}$ 22.2 %[2,4,5]
Global background	additive signal from a Gaussian distribution	fixed mean/SD derived from experimental data	not critical[6]
Local background	additive signal from fractal clouds	signal/background ratio	mean signal/background ratio $<$ 25

[1] Pearson correlation $<$ 0.95.

[2] Percent of spot radius relative to the mean spot distance.

[3] For VisualGrid and FA; AIDA did not become critical for the parameter range used for the simulations in this study.

[4] Only analysed with FA.

[5] For radius $\geq$ 0.228 mm the automatic gridfind failed.

[6] Not critical for global background noise that is comparable to our experimental reference data.

section A.3.1 for the experimental setup); the experimental images were analyzed with the in-house developed image analysis FA (see section A.3.1) and medians and the coefficients of variation (CV = standard deviation/mean) were calculated from the replicates of each clone. These data were used as the experimental reference. Fig. 2.17 shows the distributions of these medians and CVs. If reproducibility is perfect, the CV is 0, if it is poor the CV tends to higher values. The CVs of the raw data are most frequently in the interval between 0.4 and 0.5 (Fig. 2.17B). These values are fairly high since a CV of 0.5 for example means that nearly 50 % of the measurement is caused by error. However, it shall rather be an upper bound for initial data reproducibility. Only then error parameters can be identified more clearly. In published studies, the CV is in the range of 10 %–25 % (e.g. Herwig et al., 2001; Salin et al., 2002) since raw data undergoes intensive data normalization and calibration. The signals for the treatment target hybridization were derived from the medians of the experimental reference signals by upregulating 5 000 clones (35.2 % of all clones) randomly. The coefficients of these upregulations—the expression ratios—are uniformly distributed between 1 and 10.

The signals of the other 9 208 clones remained unchanged. Both signal sets consist of values for the 14 208 clones that were screened for differentially expressed genes. The input signal intensity for the spots corresponding to the constant *Arabidopsis thaliana* cDNA clones of the experimental reference was always the same. For the expression analysis, six images of filter hybridization experiments were simulated for both signal sets, respectively. Signal intensity variations as described in the following paragraph and local background noise variations (see below) were carried out for each filter. The spotting order was identical with the experimental reference.

2.3.3 Simulation Model

2.3.3.1 Generation of Signal Intensities

Schuchhardt et al. (2000) have shown that a strong correlation exists for spot intensities spotted by the same pin. Spots in the same block are spotted by the same pin. Clones that are spotted in different blocks are spotted by different pins. Thus the amount of material that is transfered to the array varies from pin to pin, and this relative pin specific variation can be described for the 384 pins of a gadget by the following pin distribution $P(Y)$

$$P(Y) = N(1, \sigma_1^2); \quad \sigma_1 = 0.43. \tag{2.22}$$

Here $N(1, \sigma_1^2)$ denotes a Gaussian normal distribution with mean 1 and variance σ_1^2. The standard deviation, σ_1, was derived from experimental data: Clones with identical 384-well microtiter plate positions were spotted by the same pin. In the experimental reference, *A. thaliana* cDNA of identical amplicons was spotted in each block as a control. Based on this information the mean CV over all pins was calculated and used as σ_1.

On one filter the signal distribution $P(X_{ij})$ of replicates is defined as follows

$$P(X_{ij}) = N(y_i \cdot z_j, (y_i \cdot z_j \cdot \sigma_2)^2); \quad \sigma_2 = 0.2 \tag{2.23}$$

with $\begin{matrix} i \in \mathbb{N}; & i \in [1, w] \\ j \in \mathbb{N}; & j \in [1, m] \end{matrix}$ z_j is the mean signal for clone j taken from the median signal distribution of experimental data (cf. Fig. 2.17), y_i denotes the pin dependent factor for pin i derived from the distribution, $P(Y)$. For the simulations presented here the number of pins is $w = 384$ and the number of clones is $m = 14\,208$. Using the duplicate correlation (0.8) of the constant experimental *A. thaliana* clone signals and σ_1 one can calculate $\sigma_2 = 0.2$, because they are associated with each other (M. Steinfath, pers. comm., proof is not shown). Thus σ_2 is the CV for identical PCR-products that were spotted by the same pin.

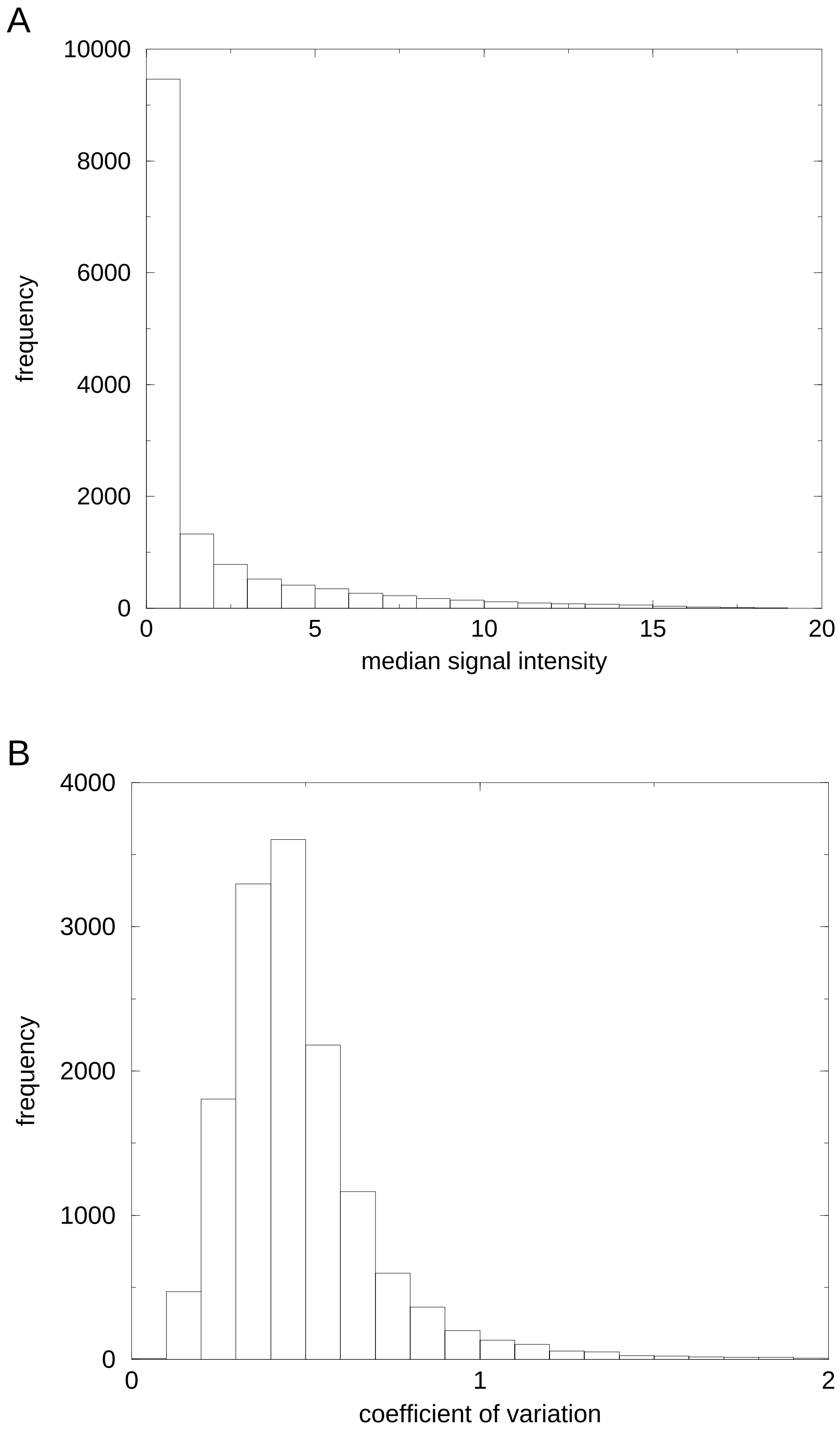


Figure 2.17: Experimental reference for simulation data. Distribution of the hybridization signals used as experimental reference. (**A**) Histogram of medians of 14 208 clones from 12 replicates each; (**B**) Histogram of coefficients of variation.

2.3.3.2 Filter Model

The simulated images are generated by an intensity function, which yields an intensity value for each pixel k. The presented model is based on empirical assumptions. It is given by a continuous function of the position $\mathbf{r}$ on the filter, $I(\mathbf{r})$, as follows:

$$I(\mathbf{r}) = \sum_j A_j f(|\mathbf{r} - \mathbf{r}_j|) + g(\mathbf{r}) + \epsilon \tag{2.24}$$

where A_j is the given spot intensity, g is a function that describes the local and global background, ϵ denotes a stochastic perturbation, and $|\mathbf{r} - \mathbf{r}_j|$ is the Euclidean distance to the center of spot j. The nine spot centers closest to $\mathbf{r}$ are considered, due to the fact, that the pixelized spot shape is given by a square 19×19 pixel matrix and the usual distance between two spot centers is 7.89 pixel for the image resolution used in this paper (0.114 mm/pixel). Here $f(|\mathbf{r} - \mathbf{r}_j|)$ is a spot shape distribution which describes the spot shape (see below). The pixel intensity $\tilde{I}(k)$ is given by

$$\tilde{I}(k) = \left[\frac{I(\mathbf{r}_k) * 2^N}{\max_\mathbf{r} I(\mathbf{r})} \right] \tag{2.25}$$

with $N = 16$ for a 16 bit image. $\mathbf{r}_k$ is the center of the pixel k. The square brackets denotes the integer function, that returns the largest integer less than or equal to the value in brackets.

The spot intensities A_j are taken from a real experiment (see above, intensity distribution see Fig. 2.17). To determine the location $\mathbf{r}_j$ of the spots I assume that the probes are spotted approximately in an orthogonal grid.

2.3.3.3 Local Distortions

Local distortions of the spots are considered. Due to the experimental procedure two different spot distortions are introduced: spot shifting and pin shifting. Both of them are modeled by randomly Gaussian distributed shifting of the spot-centers relative to their theoretical spot-centers. For spot shifting the distortions are independent for each spot; for pin shifting they are equal for all spots of one block of 5×5 spots, because they were spotted by the same pin.

2.3.3.4 Spot Shape

Due to the experimental procedure of the array preparation, the array surface type, and the nature of the fixed DNA material, the spot shapes are different. Here I introduced three distribution models of spot shapes that are based on experimental evidence:

(a) a normalized two-dimensional Gaussian distribution with a given SD (σ)

$$f(|\mathbf{r}-\mathbf{r}_j|) = \frac{1}{2\pi\sigma^2}e^{-\frac{(\mathbf{r}-\mathbf{r}_j)^2}{2\sigma^2}}, \tag{2.26}$$

(b) a normalized two-dimensional Gaussian distribution with a given SD (σ_1) of which another concentric Gaussian-distribution (SD = σ_2) with a scaling-factor $S \in (0,1)$ is subtracted. The resulting spot resembles a crater like spot shape. The derivation of the equation is shown in Appendix A.3.3.

$$f(|\mathbf{r}-\mathbf{r}_j|) = \left(\frac{1}{2\pi\sigma_1^2}e^{-\frac{(\mathbf{r}-\mathbf{r}_j)^2}{2\sigma_1^2}} - S\frac{1}{2\pi\sigma_2^2}e^{-\frac{(\mathbf{r}-\mathbf{r}_j)^2}{2\sigma_2^2}}\right) \times (1-S)^{-1}, \tag{2.27}$$

(c) a normalized cylindric distributed shape with a given radius d that forms a plateau-like spot:

$$f(|\mathbf{r}-\mathbf{r}_j|) = \begin{cases} \frac{1}{\pi d^2}, & \text{if} \quad |\mathbf{r}-\mathbf{r}_j| \leq d \\ 0, & \text{if} \quad |\mathbf{r}-\mathbf{r}_j| > d. \end{cases} \tag{2.28}$$

These spot models were used because they are commonly observable with spotted array data on nylon and glass supports respectively and are frequently assumed as quantification models by image analysis programs. More irregular spot shapes that do not have a common spot distribution can also be observed (e.g., Jain et al., 2002), but are not considered here.

2.3.3.5 Background Noise

Two different sources of background noise can be distinguished: a global background due to the scanner noise or filter surface, and a local background due to inhomogeneous hybridization to the filter that looks like smear.

Global background noise. The global background is described by a randomly Gaussian distributed noise that is equal for the whole filter. It can be varied by its mean and SD.

Local background noise. As a model for the local background, fractal clouds as described in Saupe (1988) are used. They are generated with the *midpoint displacement method* with a fractal dimension of 0.4 and then scaled to a given minimum/maximum-range, which defines the intensity level of this background. The model was chosen for local background, because the intensity level of a given pixel depends on its neighbors. This results in images that look quite the same as the background of experimental images. By the use of a pseudo random number generator, reproducible fractals were created.

2.3.4 Data Evaluation and Quality Measurement

2.3.4.1 Image Analysis

To illustrate the power of using simulated data for the judgment of image analysis software, the following programs were used: (1.) FA, which is a fully automated image analysis software—no manual effort for the positioning of the grid is necessary, (2.) AIDA, which needs some manual interaction for the positioning of the grid, and (3.) Visual Grid, for which the whole grid has to be adapted manually (see also section A.3.2). These programs have been chosen, because they are frequently used at our institute and have already been utilized intensively for image analysis (FA: Steinfath et al., 2001; Visual Grid: Herwig et al., 2001). Furthermore, they are representative for the different levels of automation of image analyses.

2.3.4.2 Evaluation of Gridfind and Quantification Quality

The following two steps are essential for the analysis of hybridization images: *gridfind* and *quantification*. First the gridfind has to locate the exact positions of the spots and then the signal intensities are assigned to each spot by the quantification. For instance, the image analysis FA does a Gaussian spot shape fit for quantification (Steinfath et al., 2001). The performance of the different image analysis programs are tested by the following quality parameters:

1. The mean distance between simulated and calculated spot centers. Here, the simulated spot center refers to the exact position of the spot center that was used for the simulation. The calculated spot center refers to the spot center that was determined by the image analysis software.

2. The Pearson correlation between simulated and calculated intensities. The simulated intensity refers to the intensity value used for the simulation and the calculated intensity is the intensity value determined by the image analysis software.

The first parameter measures the quality of the gridfind. The second is a measure for the quality of the whole image processing.

2.3.4.3 Statistical Evaluation of Differential Expression

For testing statistical significance of differential expression we calculated P values according to the Welch test (Welch, 1947). This test is an unpaired t-test. It assumes that the two samples ("treatment" and "control") are distributed according to Gaussian distributions with means, $\mu_{\text{treatment}}$ and μ_{control} respectively, and judges the hypothesis whether $\mu_{\text{treatment}} = \mu_{\text{control}}$. Here, in contrast to Student's t-test, it is not assumed that both sample distributions have the same variance. The test statistic, T, has the form

$$T = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}. \tag{2.29}$$

Here, $\overline{x}$ and $\overline{y}$ denote the sample means, S_x^2 and S_y^2 denote the sample variances and n and m are the respective sizes of the treatment and the control sample. High and low values of the test statistic then indicate significantly different sample means. This test has been applied to differential expression analysis of array data in several studies, for example Herwig et al. (2001) and Dudoit et al. (2002).

The quality of an expression profile analysis based on array data is highly dependent on the number of repeated sample measurements, and of the array preparation, hybridization and signal quantification procedure. The latter can be optimized either experimentally by improving array preparation and hybridization, or computationally by employing better algorithms for the image analysis software, such that can deal with preparation errors. The improvement of each method is limited. Major critical parameters are local distortions of the spots, variations of the spot shape and outshining effects due to neighbor spots or massive background noise. These parameters have been analyzed in this thesis (see Tab. 2.1).

In the following series of images are presented for which only one parameter was changed, respectively.

2.3.5 Simulation of Local Distortions

For the following simulation it was assumed the spots to have constant Gaussian shape without background noise. Thus only the effects of local distortions are tested. Fig. 2.18 and 2.19 show the influence of spot-shifts on the gridfind and quantification.

2.3.5.1 Spot Shifting

Spot shifting was simulated with SDs between 0 and 0.342 mm from its ideal positions (Fig. 2.18). The mean distance between adjacent spot centers was 0.9 mm. This parameter became critical (correlation < 0.95) for SDs in the range of 0.15–0.2 mm concering the quantification done with any of the three programs (Fig. 2.18B). The quantification is of course influenced by the quality of the gridfind. This means that the critical range in terms of spot shifting is about a fifth of the distance of adjacent spot centers.

In Fig. 2.18C only the quality of the gridfind is judged. The error given by the mean distance between the calculated spot center determined by the image analysis software and

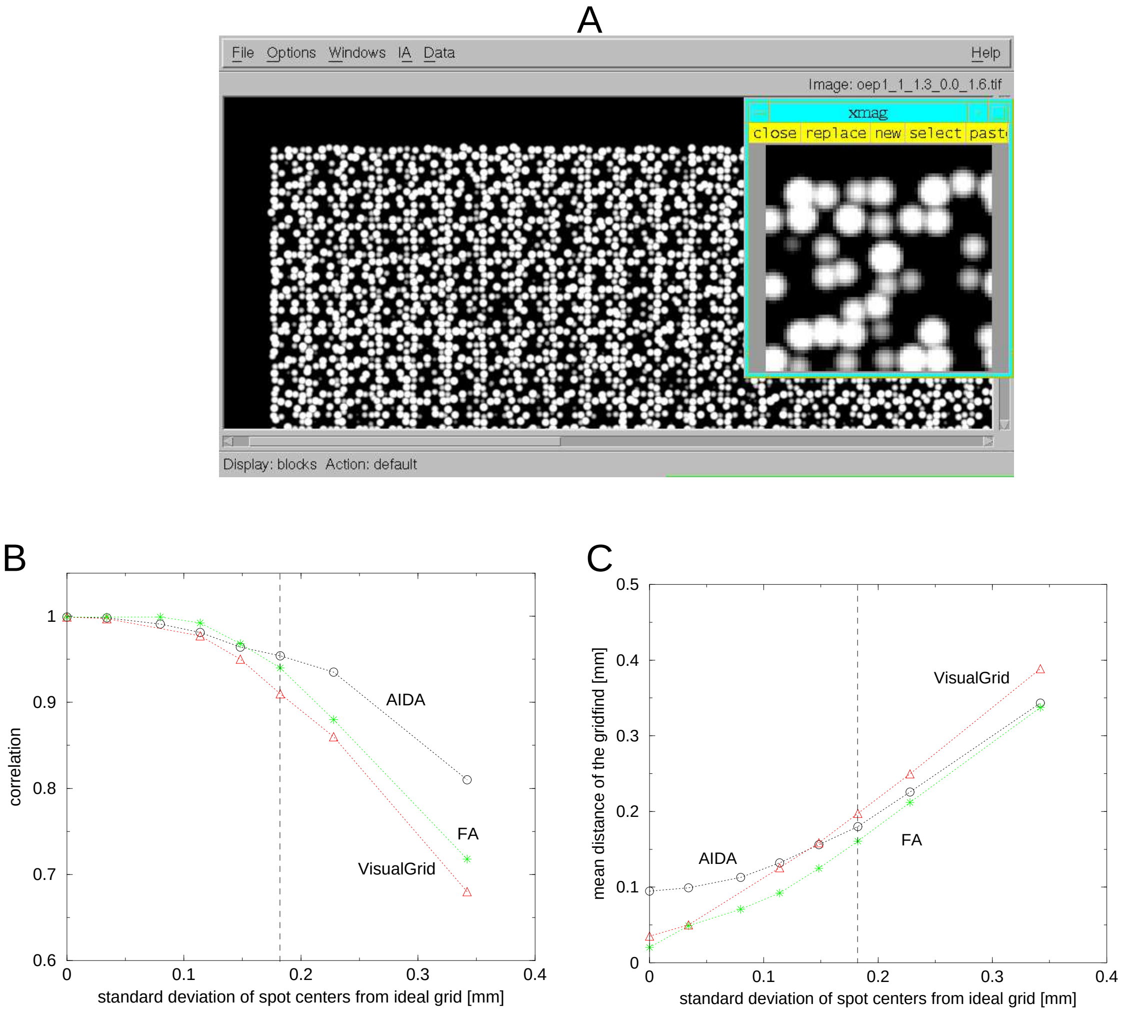


Figure 2.18: Spot shifting. Every spot was shifted randomly relative to the ideal grid position by a Gaussian distributed distance with a given standard deviation σ. (**A**) A simulated image, $\sigma = 0.1824$ mm. In (**B**) the pearson correlation between simulated and calculated intensities is plotted versus the standard deviation of the spot centers from their ideal grid nodes. In (**C**) the mean distance between the calculated and the simulated spot centers is plotted versus the standard deviation of the spot centers from their ideal grid nodes. The vertical lines in (B) and (C) correspond to the image in (A). In (B) and (C) each point in the plot is determined by a single analysis of a simulated image, respectively.

its simulated center is relatively linear to its perturbation for any of the tested programs. The low quality for AIDA for small perturbations is due to a missing sub-pixel precision. This means, that if e.g., the simulated spot center is not identical with the center of a pixel. The output-result from AIDA lacks this sub-pixel precision.

2.3.5.2 Pin Shifting

The error due to pin variations is a systematic error for all spots in the same block, because they were spotted by the same pin (Fig. 2.19A). Perturbations with SDs between 0 and 0.2 mm were simulated. This error became critical (correlation < 0.95) for SDs of the pin shifting greater than 0.12 mm for Visual Grid and greater than 0.167 mm for FA. The error of the gridfind was linear to its perturbation (Fig. 2.19C). Here again the low quality for AIDA for small perturbations is due to the missing sub-pixel precision.

Fig. 2.20 shows the distribution of block center shifts measured for experimental data (the block centers were manually determined with Visual Grid). For the results mentioned above this means that the error due to pin shifting is never in the critical area for the majority of blocks. Nonetheless, strongly depending on the used devices (e.g., spotting robots), this can become a critical parameter.

2.3.6 Simulation of Different Spot Shapes

The spot shape that depends on several properties specific to spotting procedure like the spotting method, the carrier surface or the probe viscosity, was modeled as a two-dimensional Gaussian distributed shape, a crater-like shape (Fig. 2.21A–J) and a plateau shape (Fig. 2.21K). A mean SD of 0.1482 mm for a two-dimensional Gaussian distributed spot shape was handled by all three image analyses (correlation always > 0.99). Crater-like spot shapes were simulated with crater-radii ranging from 0.0285 mm to 0.285 mm (in 0.0285 mm steps; $\sigma_1 = 0.1482$ mm). To judge the influence of this parameter, the images were analyzed by FA: Up to a crater-radius of 0.1995 mm FA analyzed them without any problems (correlation always > 0.99). For crater-radii of 0.228 mm and above (Fig. 2.21H–J) FA failed due to problems during the gridfind. A third very idealized spot shape—a plateau-like spot shape—was also simulated, to see if this can be handeled by FA. Therefore, a filter with plateau spots with a radius of 0.342 mm (not overlapping with neighbor-spots; half distance between two neighbor-spots is 0.44973 mm) was simulated and has been analyzed by FA without any problems (correlation > 0.99).

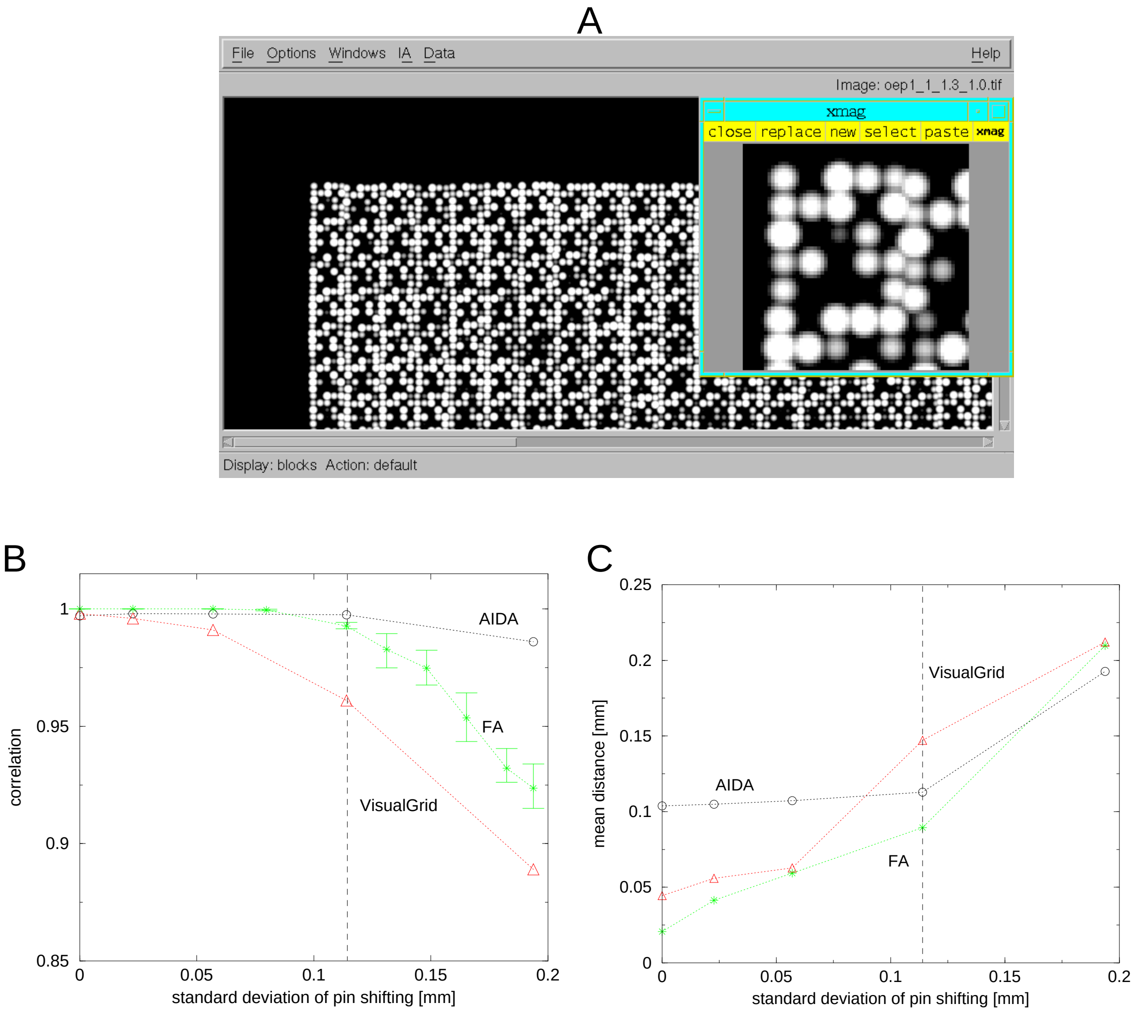


Figure 2.19: Pin shifting. Every block was shifted randomly relative to its ideal position by a Gaussian distributed distance with a given standard deviation σ. (**A**) simulated image, $\sigma = 0.114$ mm, (**B**) Pearson correlation of simulated and calculated intensities dependent on the standard deviation of the block centers from their ideal positions (for AIDA and Visual Grid each data point is determined by a single analysis of a simulated image and for FA three different images have been analyzed for each σ, the asterisk depicts the mean and the error bars show the minimum and maximum value of the three repetitions), (**C**) mean distance between the calculated and simulated spot centers dependent on the standard deviation of the block centers from their ideal positions (each data point is determined by a single analysis of a simulated image). The vertical lines in (B) and (C) correspond to the simulated image in (A).

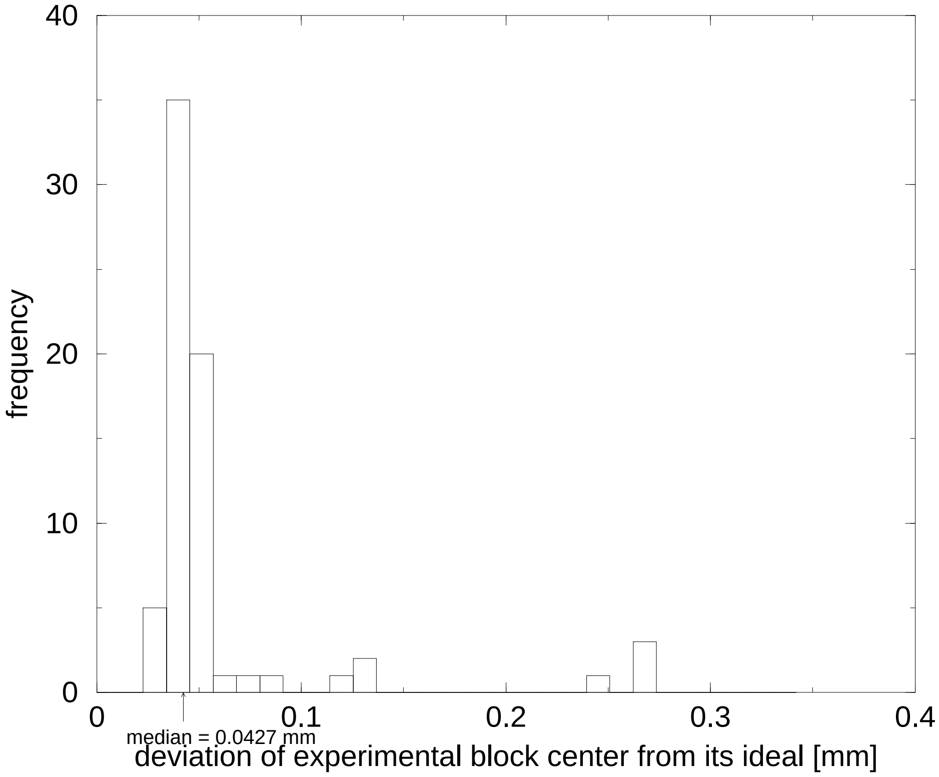


Figure 2.20: Experimental block center deviation. Histogram of the distance of experimental block centers from their ideal block centers (computed from 12 experimental filter-images each containing 48×48 blocks with 5×5 spots respectively). Block positions were manually tagged by the use of Visual Grid and distances to the ideal grid—given by field corners—were calculated.

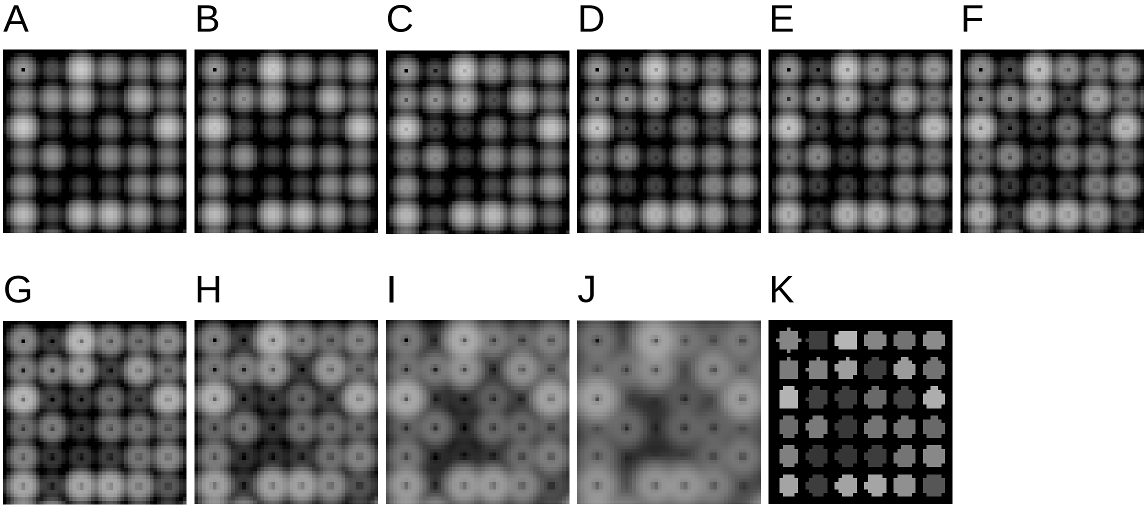


Figure 2.21: Spot shape examples. (**A**–**J**) are examples of simulated crater spot shapes with rim radii between 0.0285 mm and 0.285 mm in 0.0285 mm steps. (**K**) is an example of a plateau spot shape (radius = 0.342 mm).

2.3.7 Simulation of Background Noise

In the following all images were assumed to have constant Gaussian spot-shapes and all spot centers are located at the ideal grid nodes. Thus the gridfind has only to cope with the background noise.

2.3.7.1 Global Background Noise

From the (non-spotted) border area of an experimental filter image with a 16 bit depth, the noise level was found to be about 16000 with a standard deviation of about 4000; the distribution is similar to Gaussian (data not shown).

The simulated image shown in Fig. 2.22A has Gaussian background noise with $\mu = 16000$ and $\sigma = 4000$. The detection of the grid was nearly perfect for all image analysis programs for this image. The correlations between input and output intensities were always higher than 0.99. Hence a realistic global background noise as given by the experimental reference does not influence the quantification of the programs.

2.3.7.2 Local background noise

As a model for the local background, fractal clouds as described in Saupe (1988) were used (Fig. 2.22B).

Fig. 2.23 shows the effect of local background-noise on the image analysis. For mean signal/background ratios above 25 this error did not become critical, as far as any of the three programs are concerned. Below a ratio of 20, correlation decreases rapidly, especially for AIDA. Correlations for Visual Grid and FA decrease significantly for mean signal/background ratios below 13. At this point the signal/background ratio becomes critical for all programs. Thus it was chosen for a further statistical test series (see below).

2.3.8 Simulation of the Influence of Background Noise on the Expression Analysis

I investigated into the influence of local background noise on the quality of the expression analysis with varying numbers of repetitions. The significance of differentially expressed genes was judged by the use of the Welch test as described in Herwig et al. (2001).

A series of six images with variations in signal intensities due to replicated spotting of duplicates, and with a varying transfer quality for different pins as described in section 2.3.3.1 was simulated. Furthermore, different local backgrounds with intensities scaled in the same way as given for the mean signal/background ratio of 13 as described in section 2.3.7.2 were

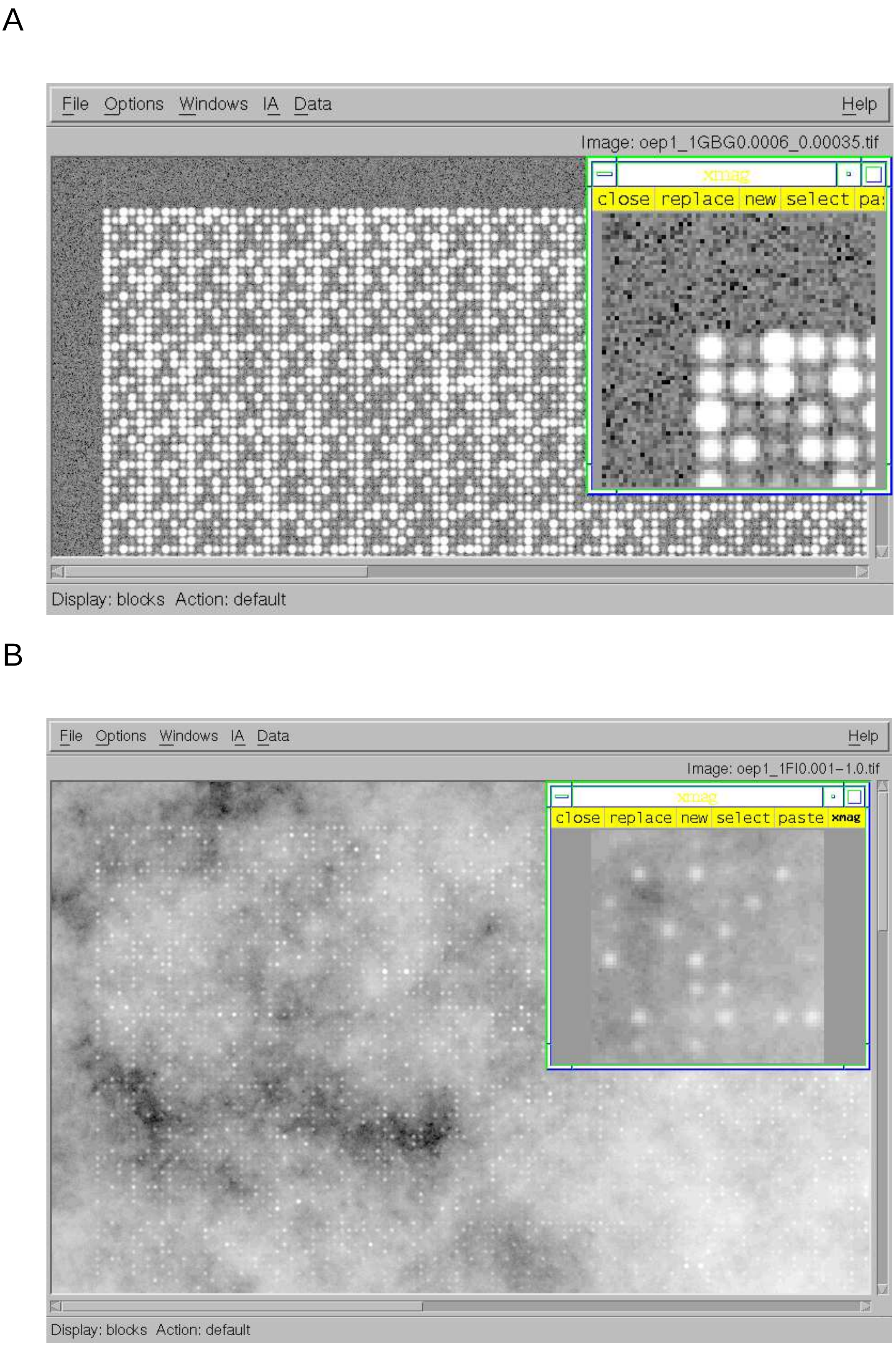

A
File
Options
Windows
IA
Data
Help
Image: oep1_1GBG0.0006_0.00035.tif
xmag
close
replace
new
select
pa:
Display: blocks Action: default
B
File
Options
Windows
IA
Data
Help
Image: oep1_1FI0.001−1.0.tif
xmag
close
replace
new
select
paste
xmag
Display: blocks Action: default

Figure 2.22: Background noise examples. Examples for filter images with simulated global (**A**) and local (**B**) background noise.

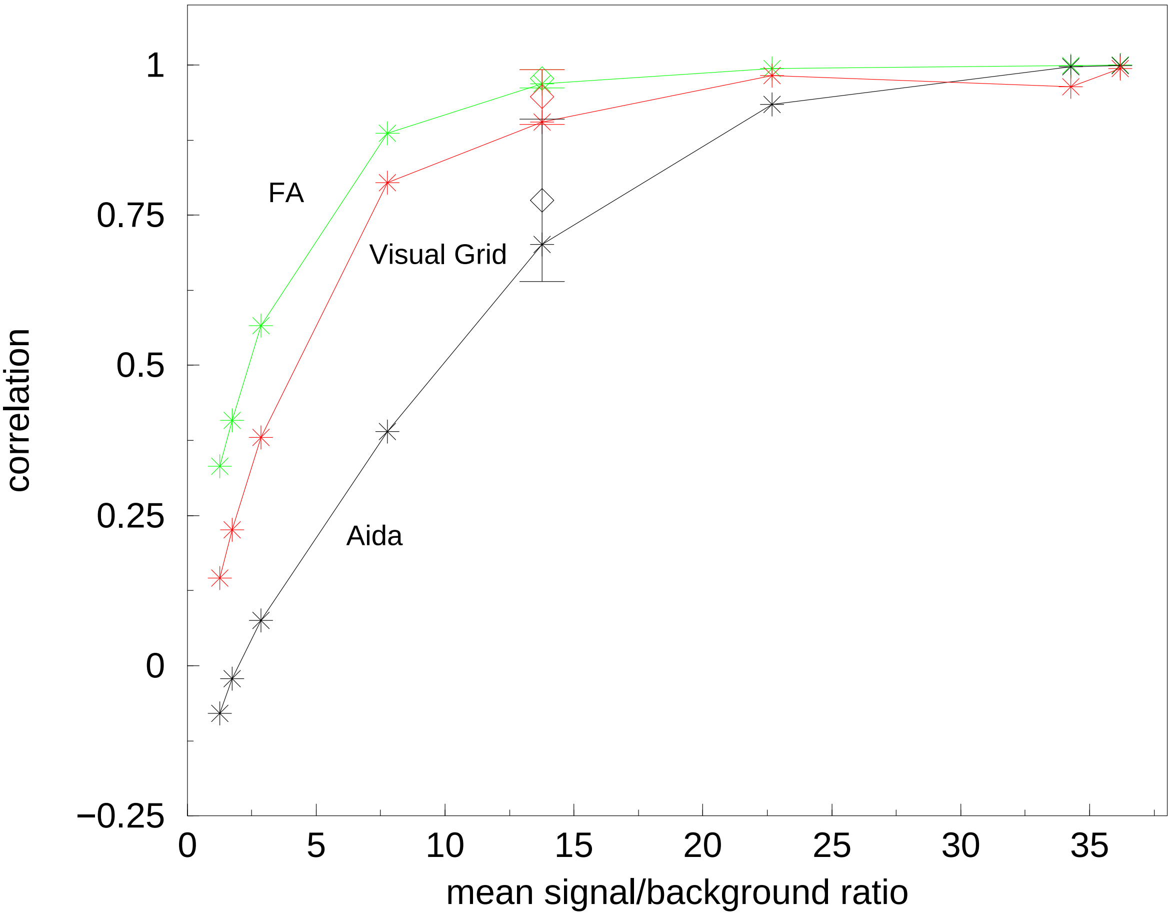


Figure 2.23: Correlation for local background noise between simulated and calculated intensities. Pearson correlation between simulated and calculated intensities depending on the intensity-level of the fractal background given by the mean of all signal/background ratios over all spots. Each data point (asterisk) corresponds to the results of one image analysis. The used fractal background image was always identical except for the signal/background ratio of 13. For this ratio, 7 different fractal background images were simulated; correlation mean μ (diamond) and standard deviations (error bars representing the interval $\mu \pm \sigma$) were calculated.

added. This was carried out for a control set with 14 208 different test clones and for a test set. For the latter signal intensities of 5 000 clones were up-regulated with factors varying between 1 and 10. Images were analyzed by three image processing programs, namely FA, AIDA, and Visual Grid. The source signal sets used for the individual image simulations as well as the analyzed data were used for the statistical significance test. The test was carried out for two, four, and six images of the control and test series, respectively. This corresponds to samples with four, eight, and twelve signals per clone and series. Results are depicted in Fig. 2.24. The rate of false positive clones is always low (false positive rate < 0.02). For input data (Fig. 2.24A) with expression ratios below 1.45, merely 42 % of the regulated clones (sample size 12) could be identified (P value < 0.01). For expression ratios above 1.45 and sample size 12, almost all regulated clones could be identified. For ratios above 1.9, a sample size of 8 was sufficient for significant identification of nearly all regulated clones. For a sample size of 4 even with ratios between 9.55 and 10.0 only 93 % of the regulated clones could be identified, while for sample size 8 and 12, 98.5 % were found. After image analysis the number of identified regulated clones decreased significantly. With the image analysis FA and sample size of 12, more than 90 % significant clones could be found for expression ratios above 1.9 (Fig. 2.24B). AIDA (Fig. 2.24C) and Visual Grid (Fig. 2.24D) needed a sample size of 12 and ratios above 3.7 to detect as many. Especially for expression ratios between 1.45 and 1.9, with FA (sample size 12) 89 % of the regulated clones could be identified, while AIDA identified only 67 % and Visual Grid 61 %. However, expression ratios of smaller than 2 seem to be critical for this kind of expression analysis. For expression ratios above 2 the differences between sample size 8 and 12 are relatively small as compared to sample size 4.

Fig. 2.24E shows a comparison of the CVs for sample size 12 of the input data signals and of the signals quantified by the three different image processing programs. The medians of the CVs increase in the following order: input data (0.19), FA (0.21), AIDA (0.29), Visual Grid (0.34). This result shows that data reproducibility increases with the level of automation of the image analysis programs.

2.3.8.1 Summary of the Findings

Complex hybridization experiments are based on a data production pipeline that incorporates a significant amount of error parameters. Here I presented a simulation environment to judge the influence of different parameters, like spot shapes, spot positions, and local and global background noise on the subsequent data analysis, such as image analysis and the detection of differentially expressed genes. Image analysis can be classified by manual, semiautomated, and automated procedures. While manual methods rely on a strong supervision by

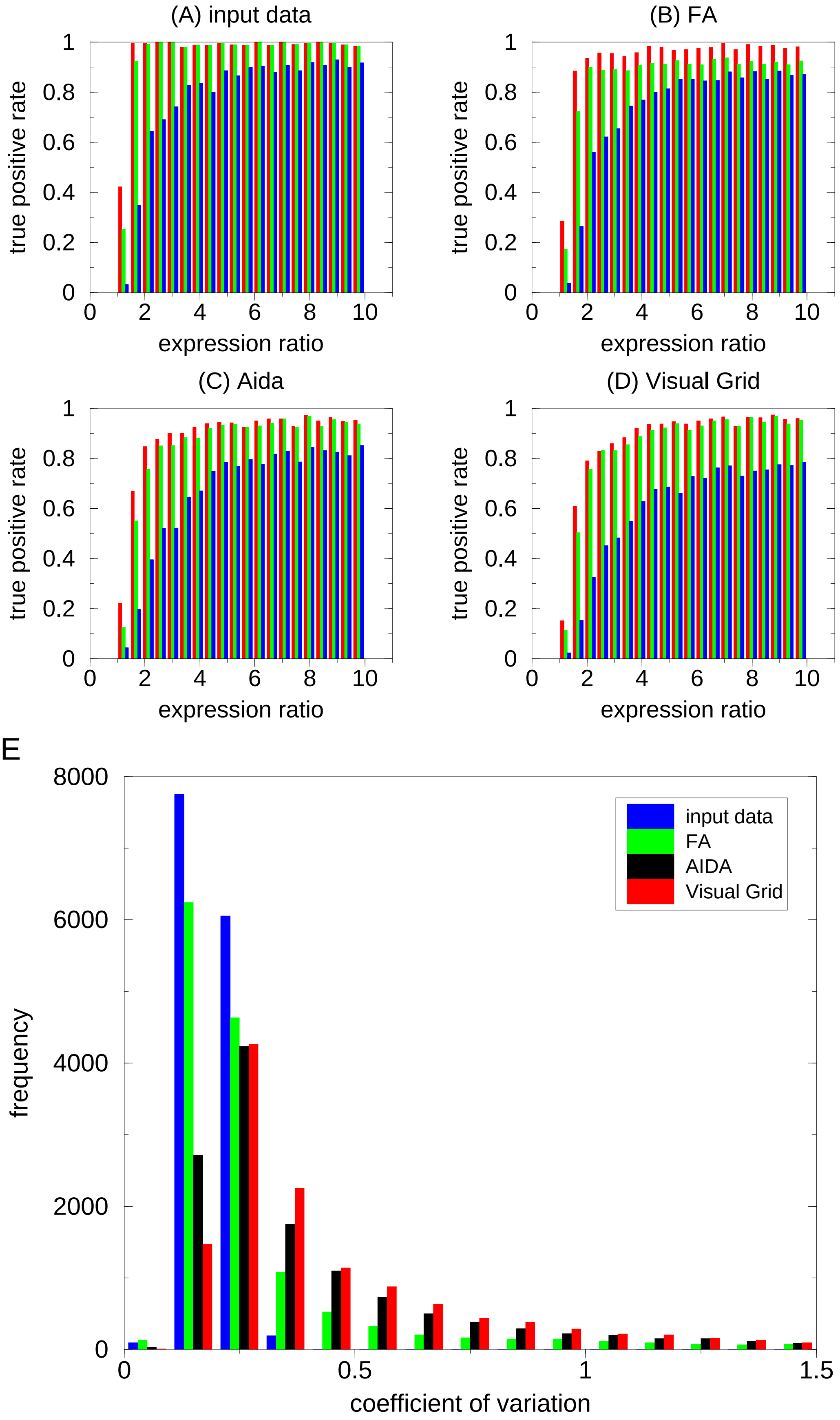

(A) input data
(B) FA
(C) Aida
(D) Visual Grid
true positive rate
expression ratio
0
0.2
0.4
0.6
0.8
1
0
2
4
6
8
10
E
frequency
coefficient of variation
0
2000
4000
6000
8000
0
0.5
1
1.5
input data
FA
AIDA
Visual Grid

Figure 2.24: (on the previous page) **Results of statistical tests for simulated fold-changes.** True positive rates of detected simulated fold-changes (P value < 0.01) as given by the Welch test. For all test results, the false positive rate is below 0.02. (Histogram intervals have a width of 0.45. The absolute number of regulated clones per interval ranges between 217 and 289.) (**A**) Simulated signals without image analysis (input for the image simulation); and after image analysis of the simulated images with FA (**B**), AIDA (**C**), Visual Grid (**D**). For all expression ratio intervals results for 12 (red), 8 (green) and 4 (blue) repetitions are given. (**E**) Histogram of the distribution of the CVs for sample size 12; The medians of the CVs are the following: input data: 0.19, FA: 0.21, AIDA: 0.29, Visual Grid: 0.34.

the user and requires some initial guess, e.g., on the spot positions, semiautomated methods require much less interaction, but still need prior information (e.g., definition of the spotted area). Automated methods try to find the spot grid without user interaction. The simulation studies have shown that the data reproducibility increases with the grade of automation of the software. However for noisy hybridization images that show very irregular structures, manual methods might be the best choice. My results show that the simulation tool is a valuable resource for the identification and the rating of error sources arising from hybridization experiments. The simulated sets can be used as benchmark tests for new data analysis modules such as image analyses coming up in the course of gene expression data analysis or comparable array based methods.

3 Discussion

Modeling and simulation techniques are valuable tools for the understanding of complex systems. In the course of my thesis I have applied modeling strategies to biotechnological laboratory methods and biological systems. During the last years, high throughput technologies are more and more frequently used in biological research. In particular, array-based gene expression analysis became an important key technology for genome and transcriptome analysis. Such array-based analysis make use of complex production pipelines that incorporate a significant amount of error parameters. In the previous chapter (Section 2.3) I describe an implemented model for the simulation of DNA-array experiments that was used to judge the influence of critical parameters on subsequent image analysis and differential expression analysis. Parts of the model have already been used for additional research by other scientists. This is discussed in Section 3.3.

Application of modeling approaches to biological systems became very popular in recent years in the course of systems biology. Since this is a very young research area, there is still a demand for appropriate computational tools. As biological systems are composed of complex interaction networks consisting of thousands of individual molecules each with different functions, the demand for integrative systems biology platforms that can cope with such large and complex interaction networks is high. One goal of my thesis was to identify and implement appropriate methods for the development and simulation of cellular reaction networks. Therefore, I have developed the modeling and simulation software application PyBioS that is available through the Web. In Section 3.1 I discuss the functionalities of PyBioS further improvements.

Moreover, PyBioS was used for the modeling of signal transduction pathways and subsequent gene-regulatory target genes related to somitogenesis. Since biological reaction networks are highly interwoven and established mechanism often are reused the developed models of Notch, Wnt, and FGF signaling including their regulation of gene targets is also of high interest to areas of research. For instance Notch, Wnt, and receptor tyrosine kinase signaling pathways are also very important for many aspects of cellular processes and not at least their relevance for the onset of diseases, such as cancer (e.g. Hanahan and Weinberg, 2000).

3.1 PyBioS - a Modeling and Simulation Platform for Cellular Reaction Networks

Compared to other systems biology software applications, PyBioS has some unique features that are particularly useful for the automated or semi-automated model development or the visualization of reaction networks along with simulated time course data. These features makes PyBioS also applicable the work with large reaction networks.

Another feature that distinguishes PyBioS also from many other system biology application is its Web-based user interface. Lee et al. (2008) compared five different Web-based simulation tools including PyBioS. Advantages of Web-based simulation platforms are, for example, that they operate through a Web browser and are, therefore, easily accessible on different platforms. Moreover, it is not necessary to install a local copy of the software as well as subsequent upgrades or bug fixes. However, Web-based applications do suffer from a significant disadvantage in speed.

Another major demand from a modeling tool for the development and representation of models of biological systems is the support for the visualization of the reaction network. Graphical representations of reaction networks prove as very helpful tools for the work in systems biology. The graphical representation of a reaction system is not only helpful during the design of a new model and as a representation of the model topology, it is also helpful for the analysis and interpretation for instance of simulation results. Traditionally, diagrams of interacting enzymes and compounds have been written in an informal manner of simple unconstrained shapes and arrows. Several diagrammatic notations have been proposed for the graphical representation (e.g., Kohn, 1999; Pirson et al., 2000; Kitano, 2003; Kitano et al., 2005; Moodie et al., 2006) As a consequence of the different proposals the Systems Biology Graphical Notation (SBGN) has been set up recently. It provides a common graphical notation for the representation of biochemical and cellular reaction networks. SBGN defines a comprehensive set of symbols, with precise semantics, together with detailed syntactic rules defining their usage. Furthermore, SBGN defines how such graphical information is represented in a machine-readable form, to ensure its proper storage, exchange, and reproduction of the graphical representation.

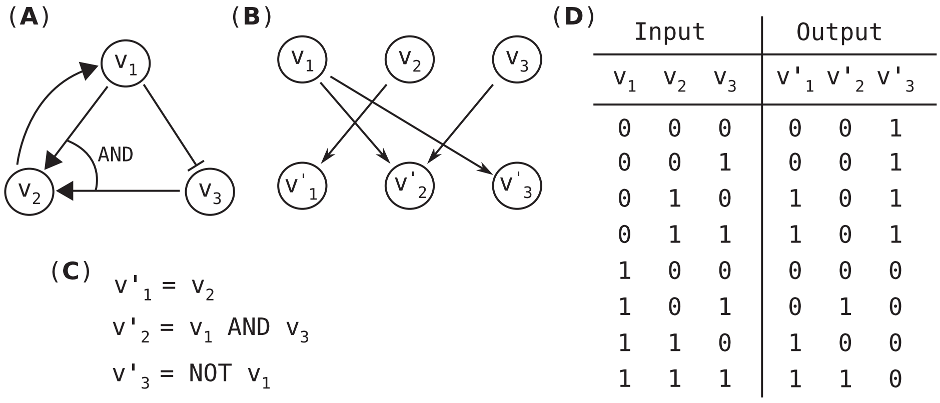
SBGN defines three different diagram types: (1) State Transition diagrams that are depicting all molecular interactions taking place, (2) Activity Flow diagrams that are representing only the flux of information going from one entity to another, and (3) Entity Relationship diagrams that are representing the relationships between different molecular species. In a State Transition diagram, each node represents a given state of a species, and therefore a given species may appear multiple times. State Transition diagrams are suitable for following the temporal process of interactions. A drawback of State Transition diagrams, however, is that the representation of each individual state of a species results quickly in very large diagram and due to this it becomes difficult to understand what interactions actually exist for the species in question. In such a case an Entity Relation diagram is more suitable. In an Entity Relation diagram a biological entity appears only once.

SBGN defines several kinds of symbols, whereas two types of symbols are distinguished: nodes and arcs. There are different kinds of nodes defined. Reacting state or entity nodes represent, e.g., macromolecules, such as protein, RNA, DNA, polysaccharide, or simple chemicals, such as a radical, an ion or a small molecule. Container nodes are defined for the representation of a complex, compartment or module. Different transition nodes are defined for the representation of transitions like biochemical reactions, associations, like protein-complex formation, or dissociations, like the dissociation of a protein complex. The influence of a node onto another is visualized by different types of arcs representing, e.g., consumption, production, modulation, stimulation, catalysis, inhibition or trigger effect. Not all node and arc symbols are defined for each of the three diagram types. A detailed description of the different nodes, arcs and the syntax of their usage by the different diagram types is given in the specification of SBGN (see http://sbgn.org/). The SBGN notation defines a more complex representation than it is provided by PyBioS at the moment.

Besides graphical aspects, also modeling approach is of relevance. Different theoretical attempts have been made to describe biological systems. Deterministic approaches are based on the exact computation of changes during time. One approach that is often used, e.g., for the description of gene regulatory networks, are Boolean networks (Kauffman, 1993; Akutsu et al., 1999; de Jong, 2002). Boolean networks take into account only two states for a variable, true and false or 1 and 0. A Boolean nework is defined by a given number of binary variables and a set of Boolean rules—logical expressions that define the state of a given output variable based on a set of given input variables. An example of a Boolean network is shown in Fig. 3.1.

An extension of Boolean models are discrete models. In contrast to the two different states that are possible for Boolean models, variables of a discrete model can take a limited number of predefined discrete values.

Deterministic modeling using ordinary differential equations (ODEs) as used by PyBioS has been applied very successfully to different problem in biology. Nevertheless, modeling by differential equations ignores the stochastic nature of biology. In biochemical networks an integer number of molecules react when they collide after random times, driven by Brownian motion. One assumption for the application of ODEs is that the number of interacting molecules is very large and stochastic effects average out. For instance, this assumption applies in most cases to metabolic networks, but for the description of gene expression events



$v'_1 = v_2$

$v'_2 = v_1 \text{ AND } v_3$

$v'_3 = \text{NOT } v_1$

Input			Output		
v_1	v_2	v_3	v'_1	v'_2	v'_3
0	0	0	0	0	1
0	0	1	0	0	1
0	1	0	1	0	1
0	1	1	1	0	1
1	0	0	0	0	0
1	0	1	0	1	0
1	1	0	1	0	0
1	1	1	1	1	0

Figure 3.1: Illustration of a Boolean network (**A**) and its wiring diagram (**B**). Applying the Boolean rules (**C**) to a given input state determines a certain output state (adapted from Akutsu et al., 1999).

this might be inappropriate (Elowitz et al., 2002; Raser and O'Shea, 2004; Tang, 2008).

Stochastic kinetics of well-mixed chemical systems can be simulated using the exact methods of Gillespie (Gillespie, 1977). However, using stochastic simulations applied to large models do not scale very well. A workaround are hybrid solutions, algorithms making use of both deterministic and stochastic simulation where appropriate. The development of such approaches is still subject of current research (Kiehl et al., 2004; Griffith et al., 2006; Wilkinson, 2006).

3.1.1 Prediction in the Face of Uncertainty

Predicting effects of perturbations of complex biological systems is key to being able to solve many important problems, in particular in the case of human diseases. It is highly likely, that such predictions will have to be based on computer models that represent all relevant components of the networks involved as well as their interactions in sufficient detail and accuracy. Establishment of such models is however complicated by the fact, that relevant parameters are either completely unknown, or can at best be measured under highly artificial conditions.

Structure and behavior of any cell and any organism are determined by converting information in the genome and the environment into the phenotype through a series of molecular processes. Dysfunctions in the molecular interaction networks carrying out this process can cause severe diseases such as cancer. Curing the disease, or at least ameliorating the symptoms, often involves by itself complex disturbances in these networks, with success depending, among other factors, on the genotype of the patient. It is therefore not too surprising, that in many cases only a sometimes small fraction of patients responds to specific treatments, while many might suffer often quite severe side effects. Progress in the treatment of diseases in individual patients will therefore depend critically on being able to predict the effects of such treatments, regarding the genomic predisposition of the patient.

The capability to predict has been a main goal of science from the beginning. In contrast to the situation in many areas in physics, where it has been possible to make highly accurate predictions based on a small number of assumptions, accurate predictions of biological processes depend on the behavior of complex networks of molecular, cellular, and even organismal interactions, which have been shaped by events hundreds of millions of years ago. It is therefore quite likely that predictions in biology will have to be based to a large extent, on the detailed knowledge of the components of the networks involved, as well as their interactions. While inherently difficult to achieve, any progress in our ability to predict the behavior of these biological networks can have enormous practical consequences. Improved predictions on the response of individual patients could, for example, decide between life and death of the individuals involved, while improved predictions on the effect of drugs could very well help to revolutionize drug development, and therefore have enormous economic value.

To allow such predictions, two basic strategies have been considered: the identification of statistical correlations in the therapy response of specific *biomarkers* (e.g. transcripts, proteins, metabolites, patterns of genomic methylation, etc.), and the modeling of the disease and therapy, to represent accurately the biological processes in the individual patient. While statistical procedures have been quite successful in, e.g., predicting treatment responses, they are, however, inherently a relatively blunt instrument, only able to detect very strong correlations, which hold up across large groups, irrespective of the multiple differences between the individuals, which make up these groups. Predictive models, in contrast, can take into account the individual situation in every patient, and could therefore, in many cases, provide more reliable predictions.

The establishment of such predictive models is however complicated by the lack of information on many of the reaction kinetics needed. Information on the kinetics and kinetic parameters is either not available at all, or, at best, is based on experiments often carried out under conditions quite different from those in the living cell. Concentrations of many reactants are usually unknown, or it is simply not feasible to determine them for every individual patient. Thus, computational modeling approaches must primarily face the challenge of coping with this lack of information.

One approach to overcome this limitation could be a rigorous analysis of the model's parameter space, e.g., by sampling unknown kinetic parameters from appropriate random distributions and a subsequent statistical significance testing. Of course, such a kind of Monte Carlo-based approach requires to run thousands of simulations. This can only be performed in parallel using distributed computing like grid computing. PyBioS is already designed for such applications. Models can be developed within PyBioS and exported as model-specific software applications that no longer depend on the PyBioS system itself. This makes it possible to distribute the simulation tasks on a computer cluster. Using such a kind of *in silico* approach, one can perform experiments that might allow predictions about the effects of specific perturbations that introduced into the model.

3.1.2 Applications

The PyBioS modeling and simulation system comes already to application in several national and international research projects. A major part of the tool has been developed during the EMI-CD project supported by the European Union in its framework program 6 (FP6). Table 3.1 gives an overview of different projects PyBioS is involved in.

Table 3.1: PyBioS is developed and used in several projects supported by the European Union (EU) and the German Federal Ministry of Education and Research.

Project	Description
	Projects supported by the European Union
EMI-CD	European modelling initiative combating complex diseases
ESBIC-D	European Systems Biology Initiative combating complex diseases
EMBRACE	A European Model for Bioinformatics Research and Community Education
CARCINOGENOMICS	Development of *in vitro* test methods for identification of carciongenic substances
APO-SYS	Apoptosis Systems Biology Applied to Cancer and AIDS
	Projects supported by the BMBF
METASTEM	NMR Metabolic Profiling of the Stem Cell Niche
Mutanom	Functional characterization of mutations causing cancer
MoGLI	Systems scale analysis and modeling of Hedgehog/GLI

3.2 Modeling Biological Systems

In the past, investigations into cellular and molecular processes were done by the analysis of particular pathways, e.g., metabolic or signal transduction pathways, and the isolation and characterization of components involved in these processes. The results of these functional characterizations and analyses of many single genes are well documented in the literature, and sometimes also systematically summarized in databases.

Besides this, biological systems have features that arise from their complex interaction structure. In such systems, changes of a single component might influence several others and due to this they show a significantly different dynamic behavior. For instance, the variation of a single transcription factor might influence the expression of several of its target genes, and this results in alterations of processes these targets are involved in. Another example is given by cross-talks between different signal transduction pathways. As a consequence, interwoven networks occur that make the system much more complicated and less predictable. Thus, functions in biological systems rely on a combination of the network and the specific elements involved, and, in this way, biological systems might be better characterized as symbiotic systems (Kitano, 2002). To investigate their properties, it is necessary to consider and analyze the components in a broad context using a systems approach. For this purpose new experimental methods were developed offering tools for the analysis of different categories of the biological system. Frequently the names of the new approaches carry the suffix *omics* as is the case with genomics, proteomics, metabolomics, transcriptomics, or interactomics. In the following some of the methods utilized by the disciplines mentioned above are listed. Often they display the same methodical approach by making use of high-throughput technologies.

The prototype model of coupled Wnt, Notch, and FGF signaling, which I have implemented in the course of my thesis is not only useful for the description of somitogenesis and other developmental processes, but it can also be used for the study of disease or aging processes.

3.3 Modeling of Laboratory Methods - DNA Array Experiments

In Section 2.3 I have presented a simulation for complex hybridization experiments. This was used to judge critical experimental parameters in the light of the following data analysis. I studied critical parameter of the image analysis by the use of three different image analysis programs representing different levels of automation of the gridfinding and signal quantification. I showed that local distortions of the spot centers like non systematic spot shifting as well as systematic errors resulting in block shifting due to pin errors did not become critical for the reference experiments with the image analysis programs. Also global background noise did not become critical for the experiments studied here. Local background noise might become critical for filter experiments in some cases. Here I showed by the use of fractal clouds as background—which looks very similar to the smear in real experiments—that a mean signal/background ratio below 13 might become critical for some image analysis. However, for the automation of complex hybridizations it might be very helpful to check these parameters during the following data analysis pipeline. This can help to identifiy bad experiments more efficiently. Furthermore it might help to detect sources of error during the experimental procedure or improvements that were made. Although it is possible to get a higher quality of the results by an improvement of the experimental procedure and data analysis algorithms, it is always limited (not at least by the available resources). Furthermore variations of biological material can be expected. To cope with this limitations repetitions of the experiments are indispensable. Not at least due to the fact that array experiments are still very expensive one wants to know how many repetitions are necessary to ensure a certain quality for your expression analysis. For this purpose I did statistical analysis with 4, 8 and 12 repetitions using a Gaussian distributed noise of the input data with $\sigma_2 = 0.2$. Here the image analyses had to cope with changing local backgrounds with the same intensity level. The results of the statistical analysis indicate that for the different image analyses expression ratios below 2 become critical. The relatively poor performance for Visual Grid indicated by the distribution of the CVs is probably due to the fact that this program does no local alignment of the spot position. Since here ideal spot positions were simulated this can explain the relatively good correlation found in Fig. 2.23 for this program. But due to the manual positioning of the global grid this might become a significant source of error. AIDA and FA do local alignments for the spot positions whereby this source of noise due to manual interaction does not occur.

Automated expression analysis by chip technology will become more and more important in the future, e.g., in biology for comprehensive studies of any kind of developmental processes or in medicine for the study of genetically reasoned diseases. Therefore it is essential to have a well characterized chip technology and subsequent data analysis. This can be supported significantly by well defined models and a whole process simulation. By using well characterized radioactively labeled filter cDNA-arrays, I showed that the simulation of this biotechnological method reveals for several parameters the level when they become critical for the follow up data analysis and how this can be improved. Furthermore, the simulation environment can also be easily used for the study of cDNA arrays based on glass slides, where, e.g., background noise seems to be less critical, but distortions of spot positions and less well characterized spot shapes are more critical.

Since the simulation approach of DNA array hybridization experiments that is presented in my thesis was already published 2002 in BMC Bioinformatics (Wierling et al., 2002), its results had already impact on further research of other scientist. The models of macroarray spot shapes developed and described in this thesis (published in 2002, Wierling et al.) have been adopted by Ekstrøm et al. (2004) to fit the characteristics of microarray spot shapes more precisely.

Bibliography

ADJAYE J, HUNTRISS J, HERWIG R, BENKAHLA A, BRINK TC, WIERLING C, HULTSCHIG C, GROTH D, YASPO ML, PICTON HM, GOSDEN RG, and LEHRACH H (2005) Primary differentiation in the human blastocyst: comparative molecular portraits of inner cell mass and trophectoderm cells. *Stem Cells* **23**(10):1514–1525, doi: 10.1634/stemcells.2005-0113.

AKUTSU T, MIYANO S, and KUHARA S (1999) Identification of genetic networks from a small number of gene expression patterns under the boolean network model. *Pac Symp Biocomput* pp. 17–28.

ALBERTS B, JOHNSON A, LEWIS J, RAFF M, ROBERTS K, and WALTER P (2008) Molecular Biology of the Cell. Garland Science, fifth edition edition.

AULEHLA A and HERRMANN BG (2004) Segmentation in vertebrates: clock and gradient finally joined. *Genes Dev* **18**(17):2060–7, doi:10.1101/gad.1217404.

AULEHLA A, WEHRLE C, BRAND-SABERI B, KEMLER R, GOSSLER A, KANZLER B, and HERRMANN BG (2003) Wnt3a plays a major role in the segmentation clock controlling somitogenesis. *Dev Cell* **4**(3):395–406.

BARRETT T, TROUP DB, WILHITE SE, LEDOUX P, RUDNEV D, EVANGELISTA C, KIM IF, SOBOLEVA A, TOMASHEVSKY M, and EDGAR R (2007) Ncbi geo: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res* **35**(Database issue):D760–D765, doi:10.1093/nar/gkl887.

BELLOT F, CRUMLEY G, KAPLOW JM, SCHLESSINGER J, JAYE M, and DIONNE CA (1991) Ligand-induced transphosphorylation between different fgf receptors. *EMBO J* **10**(10):2849–2854.

BESSHO Y, MIYOSHI G, SAKATA R, and KAGEYAMA R (2001) Hes7: a bhlh-type repressor gene regulated by notch and expressed in the presomitic mesoderm. *Genes Cells* **6**(2):175–185.

BHANOT P, BRINK M, SAMOS CH, HSIEH JC, WANG Y, MACKE JP, ANDREW D, NATHANS J, and NUSSE R (1996) A new member of the frizzled family from Drosophila functions as a Wingless receptor. *Nature* **382**(6588):225–30.

BITTNER M, ET AL. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**(6795):536–40, doi:10.1038/35020115.

BLATTNER FR, ET AL. (1997) The complete genome sequence of Escherichia coli K-12. *Science* **277**(5331):1453–74.

BRAY S and FURRIOLS M (2001) Notch pathway: making sense of suppressor of hairless. *Curr Biol* **11**(6):R217–21.

BRAZMA A, ET AL. (2003) Arrayexpress–a public repository for microarray gene expression data at the ebi. *Nucleic Acids Res* **31**(1):68–71.

BRIGGS G and HALDANE J (1925) A note on the kinetics of enzyme action. *Biochem J* **19**:338–339.

BÖTTCHER RT and NIEHRS C (2005) Fibroblast growth factor signaling during early vertebrate development. *Endocr Rev* **26**(1):63–77, doi:10.1210/er.2003-0040.

CADIGAN KM and LIU YI (2006) Wnt signaling: complexity at the surface. *J Cell Sci* **119**(Pt 3):395–402, doi:10.1242/jcs.02826.

CARY MP, BADER GD, and SANDER C (2005) Pathway information for systems biology. *FEBS Lett* **579**(8):1815–20, doi:10.1016/j.febslet.2005.02.005.

CHANCE B (1943) The kinetics of the enzyme-substrate compound of peroxidase. *J Biol Chem* **151**:553–577.

CLARK MD, HENNIG S, HERWIG R, CLIFTON SW, MARRA MA, LEHRACH H, JOHNSON SL, GROUP TW, and GROUP WUGSCST (2001) An oligonucleotide fingerprint normalized and expressed sequence tag characterized zebrafish cDNA library. *Genome Res* **11**(9):1594–602.

CONLON RA, REAUME AG, and ROSSANT J (1995) Notch1 is required for the coordinate segmentation of somites. *Development* **121**(5):1533–45.

COOKE J and ZEEMAN EC (1976) A clock and wavefront model for control of the number of repeated structures during animal morphogenesis. *J Theor Biol* **58**(2):455–476.

DAILEY L, AMBROSETTI D, MANSUKHANI A, and BASILICO C (2005) Mechanisms underlying differential responses to fgf signaling. *Cytokine Growth Factor Rev* **16**(2):233–247, doi:10.1016/j.cytogfr.2005.01.007.

DAJANI R, FRASER E, ROE SM, YEO M, GOOD VM, THOMPSON V, DALE TC, and PEARL LH (2003) Structural basis for recruitment of glycogen synthase kinase 3beta to the axin-apc scaffold complex. *EMBO J* **22**(3):494–501, doi:10.1093/emboj/cdg068.

DALE JK, MAROTO M, DEQUEANT ML, MALAPERT P, MCGREW M, and POURQUIE O (2003) Periodic Notch inhibition by Lunatic Fringe underlies the chick segmentation clock. *Nature* **421**(6920):275–8, doi:10.1038/nature01244.

DE ANGELIS MH, MCINTYRE J, and GOSSLER A (1997) Maintenance of somite borders in mice requires the delta homologue dii1. *Nature* **386**(6626):717–721, doi:10.1038/386717a0.

DE JONG H (2002) Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol* **9**(1):67–103, 1066-5277 Journal Article Review Review, Academic.

DEQUÉANT ML, GLYNN E, GAUDENZ K, WAHL M, CHEN J, MUSHEGIAN A, and POURQUIÉ O (2006) A complex oscillating network of signaling genes underlies the mouse segmentation clock. *Science* **314**(5805):1595–1598, doi:10.1126/science.1133141.

DEQUÉANT ML and POURQUIÉ O (2008) Segmental patterning of the vertebrate embryonic axis. *Nat Rev Genet* **9**(5):370–382, doi:10.1038/nrg2320.

DERISI J, PENLAND L, BROWN PO, BITTNER ML, MELTZER PS, RAY M, CHEN Y, SU YA, and TRENT JM (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* **14**(4):457–60, doi:10.1038/ng1296-457.

DERISI JL, IYER VR, and BROWN PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**(5338):680–6.

DEUFLHARD P, HAIRER E, and ZUGCK J (1987) One step and extrapolation methods for differential-algebraic systems. *Num Math* **51**:501–516.

DEUFLHARD P and NOWAK U (1987) Extrapolation integrators for quasilinear implicit odes. In P Deuflhard and B Engquist (editors), Prog. Sci. Comp. 7, pp. 37–50.

DICKMEIS T, AANSTAD P, CLARK M, FISCHER N, HERWIG R, MOURRAIN P, BLADER P, ROSA F, LEHRACH H, and STRÄHLE U (2001) Identification of nodal signaling targets by array analysis of induced complex probes. *Dev Dyn* **222**(4):571–80.

DUBRULLE J, MCGREW MJ, and POURQUIÉ O (2001) Fgf signaling controls somite boundary position and regulates segmentation clock control of spatiotemporal hox gene activation. *Cell* **106**(2):219–232.

DUDOIT S, YANG Y, SPEED T, and CALLOW M (2002) Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica Sinica* **12**(1):111–139.

EDGAR R, DOMRACHEV M, and LASH AE (2002) Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Res* **30**(1):207–210.

EKSTRØM CT, BAK S, KRISTENSEN C, and RUDEMO M (2004) Spot shape modelling and data transformations for microarrays. *Bioinformatics* **20**(14):2270–8, doi:10.1093/bioinformatics/bth237.

ELKON R, ET AL. (2008) Spike–a database, visualization and analysis tool of cellular signaling pathways. *BMC Bioinformatics* **9**:110, doi:10.1186/1471-2105-9-110.

ELOWITZ MB and LEIBLER S (2000) A synthetic oscillatory network of transcriptional regulators. *Nature* **403**(6767):335–338, doi:10.1038/35002125.

ELOWITZ MB, LEVINE AJ, SIGGIA ED, and SWAIN PS (2002) Stochastic gene expression in a single cell. *Science* **297**(5584):1183–1186, doi:10.1126/science.1070919.

FRASER CM, ET AL. (1995) The minimal gene complement of Mycoplasma genitalium. *Science* **270**(5235):397–403.

FUNAHASHI A, TANIMURA N, MOROHASHI M, and KITANO H (2003) Celldesigner: a process diagram editor for gene-regulatory and biochemical networks. *Biosilico* **1**:159–162.

GALPERIN MY (2008) The molecular biology database collection: 2008 update. *Nucleic Acids Res* **36**(Database issue):D2–D4, doi:10.1093/nar/gkm1037.

GARFINKEL D, GARFINKEL L, PRING M, GREEN S, and CHANCE B (1970) Computer applications to biochemical kinetics. *Annual Review of Biochemistry* **39**:473–498.

GIBBS RA, ET AL. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**(6982):493–521.

GILBERT SF (2003) Developmental Biology. Sinauer Associates Inc., Sunderland, MA, USA.

GILLESPIE D (1977) Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* **81**(25):2340–2361.

GINKEL M, KREMLING A, NUTSCH T, REHNER R, and GILLES E (2003) Modular modeling of cellular systems with promot/diva. *Bioinformatics* **19**(9):1169–1176.

GLANSDORFF P and PRIGOGINE I (1971) Thermodynamic theory of structure, stability and fluctuation. Wiley-Interscience, London.

GOFFEAU A, ET AL. (1996) Life with 6000 genes. *Science* **274**(5287):546, 563–7.

GOLDBETER A (1991) A minimal cascade model for the mitotic oscillator involving cyclin and cdc2 kinase. *Proc Natl Acad Sci U S A* **88**(20):9107–9111.

GOLDBETER A and POURQUIÉ O (2008) Modeling the segmentation clock as a network of coupled oscillations in the notch, wnt and fgf signaling pathways. *J Theor Biol* **252**(3):574–585, doi:10.1016/j.jtbi.2008.01.006.

GOLDFARB M (1996) Functions of fibroblast growth factors in vertebrate development. *Cytokine Growth Factor Rev* **7**(4):311–325.

GRANJEAUD S, NGUYEN C, ROCHA D, LUTON R, and JORDAN BR (1996) From hybridization image to numerical values: a practical, high throughput quantification system for high density filter hybridizations. *Genet Anal* **12**(3-4):151–62.

GRESS TM, HOHEISEL JD, LENNON GG, ZEHETNER G, and LEHRACH H (1992) Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues. *Mamm Genome* **3**(11):609–19.

GRESS TM, MÜLLER-PILLASCH F, GENG M, ZIMMERHACKL F, ZEHETNER G, FRIESS H, BÜCHLER M, ADLER G, and LEHRACH H (1996) A pancreatic cancer-specific expression profile. *Oncogene* **13**(8):1819–30.

GRIFFITH M, COURTNEY T, PECCOUD J, and SANDERS WH (2006) Dynamic partitioning for hybrid simulation of the bistable hiv-1 transactivation network. *Bioinformatics* **22**(22):2782–2789, doi:10.1093/bioinformatics/btl465.

GROTH C and LARDELLI M (2002) The structure and function of vertebrate fibroblast growth factor receptor 1. *Int J Dev Biol* **46**(4):393–400.

GULDBERG C and WAAGE P (1879) Über die chemische affinität. *J Prakt Chem* **19**:69.

HADARI YR, GOTOH N, KOUHARA H, LAX I, and SCHLESSINGER J (2001) Critical role for the docking-protein frs2 alpha in fgf receptor-mediated signal transduction pathways. *Proc Natl Acad Sci U S A* **98**(15):8578–8583, doi:10.1073/pnas.161259898.

HANAHAN D and WEINBERG RA (2000) The hallmarks of cancer. *Cell* **100**(1):57–70.

HATAKEYAMA M, KIMURA S, NAKA T, KAWASAKI T, YUMOTO N, ICHIKAWA M, KIM JH, SAITO K, SAEKI M, SHIROUZU M, YOKOYAMA S, and KONAGAYA A (2003) A computational model on the modulation of mitogen-activated protein kinase (mapk) and akt pathways in heregulin-induced erbb signalling. *Biochem J* **373**(Pt 2):451–463, doi:10.1042/BJ20021824.

HEINRICH R and RAPOPORT TA (1974) A linear steady-state treatment of enzymatic chains. general properties, control and effector strength. *Eur J Biochem* **42**(1):89–95.

HENRY CA, URBAN MK, DILL KK, MERLIE JP, PAGE MF, KIMMEL CB, and AMACHER SL (2002) Two linked hairy/enhancer of split-related zebrafish genes, her1 and her7, function together to refine alternating somite boundaries. *Development* **129**(15):3693–3704.

HERMJAKOB H, ET AL. (2004) Intact: an open source molecular interaction database. *Nucleic Acids Res* **32**(Database issue):D452–D455, doi:10.1093/nar/gkh052.

HERWIG R, AANSTAD P, CLARK M, and LEHRACH H (2001) Statistical evaluation of differential expression on cDNA nylon arrays with replicated experiments. *Nucleic Acids Res* **29**(23):E117.

HERWIG R and LEHRACH H (2006) Expression profiling of drug response–from genes to pathways. *Dialogues Clin Neurosci* **8**(3):283–293.

HINDMARSH A (1983) Odepack, a systematized collection of ode solvers. In R Stepleman and et al (editors), Scientific computing, pp. 55–64, North-Holland, Amsterdam.

HIRATA H, BESSHO Y, KOKUBU H, MASAMIZU Y, YAMADA S, LEWIS J, and KAGEYAMA R (2004) Instability of Hes7 protein is crucial for the somite segmentation clock. *Nat Genet* **36**(7):750–4, doi:10.1038/ng1372.

HOFMANN M, SCHUSTER-GOSSLER K, WATABE-RUDOLPH M, AULEHLA A, HERRMANN BG, and GOSSLER A (2004) WNT signaling, in synergy with T/TBX6, controls Notch signaling by regulating Dll1 expression in the presomitic mesoderm of mouse embryos. *Genes Dev* **18**(22):2712–7, doi:10.1101/gad.1248604.

HOOPS S, SAHLE S, GAUGES R, LEE C, PAHLE J, SIMUS N, SINGHAL M, XU L, MENDES P, and KUMMER U (2006) Copasi–a complex pathway simulator. *Bioinformatics* **22**(24):3067–3074, doi:10.1093/bioinformatics/btl485.

HUANG CY and FERRELL JE (1996) Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proc Natl Acad Sci U S A* **93**(19):10078–10083.

HUBER W, VON HEYDEBRECK A, SÜLTMANN H, POUSTKA A, and VINGRON M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18 Suppl 1**:S96–104.

HUCKA M, FINNEY A, BORNSTEIN BJ, KEATING SM, SHAPIRO BE, MATTHEWS J, KOVITZ BL, SCHILSTRA MJ, FUNAHASHI A, DOYLE JC, and KITANO H (2004) Evolving a lingua franca and associated software infrastructure for computational systems biology: the systems biology markup language (sbml) project. *Syst Biol* **1**(1):41–53.

HUCKA M, FINNEY A, SAURO HM, BOLOURI H, DOYLE J, and KITANO H (2002) The erato systems biology workbench: enabling interaction and exchange between software tools for computational biology. *Pac Symp Biocomput* pp. 450–61, journal Article.

HUCKA M, ET AL. (2003) The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**(4):524–31, 1367-4803 Evaluation Studies Journal Article.

HUGHES TR, ET AL. (2000) Functional discovery via a compendium of expression profiles. *Cell* **102**(1):109–26.

HUGHES TR, ET AL. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* **19**(4):342–7, doi:10.1038/86730.

HYNNE F, DANO S, and SORENSEN PG (2001) Full-scale model of glycolysis in saccharomyces cerevisiae. *Biophys Chem* **94**(1-2):121–63.

ITOH N and ORNITZ DM (2004) Evolution of the fgf and fgfr gene families. *Trends Genet* **20**(11):563–569, doi:10.1016/j.tig.2004.08.007.

ITOH N and ORNITZ DM (2008) Functional evolutionary history of the mouse fgf gene family. *Dev Dyn* **237**(1):18–27, doi:10.1002/dvdy.21388.

IYER VR, ET AL. (1999) The transcriptional program in the response of human fibroblasts to serum. *Science* **283**(5398):83–7.

JAIN AN, TOKUYASU TA, SNIJDERS AM, SEGRAVES R, ALBERTSON DG, and PINKEL D (2002) Fully automatic quantification of microarray image data. *Genome Res* **12**(2):325–32, doi:10.1101/gr.210902.

JOHNSON DE and WILLIAMS LT (1993) Structural and functional diversity in the fgf receptor multigene family. *Adv Cancer Res* **60**:1–41.

JOSHI-TOPE G, ET AL. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* **33**(Database issue):D428–32, doi:.1093/nar/gki072.

JOUVE C, PALMEIRIM I, HENRIQUE D, BECKERS J, GOSSLER A, ISH-HOROWICZ D, and POURQUIÉ O (2000) Notch signalling is required for cyclic expression of the hairy-like gene hes1 in the presomitic mesoderm. *Development* **127**(7):1421–1429.

KACSER H and BURNS JA (1973) The control of flux. *Symp Soc Exp Biol* **27**:65–104.

KAGEYAMA R and OHTSUKA T (1999) The notch-hes pathway in mammalian neural development. *Cell Res* **9**(3):179–188, doi:10.1038/sj.cr.7290016.

KAHLEM P, ET AL. (2004) Transcript level alterations reflect gene dosage effects across multiple tissues in a mouse model of down syndrome. *Genome Res* **14**(7):1258–67, doi:10.1101/gr.1951304.

KAMBUROV A, WIERLING C, LEHRACH H, and HERWIG R (2008) Consensuspathdb–a database for integrating human functional interaction networks. *Nucleic Acids Res* doi:10.1093/nar/gkn698.

KANEHISA M, ARAKI M, GOTO S, HATTORI M, HIRAKAWA M, ITOH M, KATAYAMA T, KAWASHIMA S, OKUDA S, TOKIMATSU T, and YAMANISHI Y (2008) Kegg for linking genomes to life and the environment. *Nucleic Acids Res* **36**(Database issue):D480–D484, doi:10.1093/nar/gkm882.

KANEHISA M and GOTO S (2000) Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**(1):27–30, 0305-1048 Journal Article.

KARP PD, OUZOUNIS CA, MOORE-KOCHLACS C, GOLDOVSKY L, KAIPA P, AHRÉN D, TSOKA S, DARZENTAS N, KUNIN V, and LÓPEZ-BIGAS N (2005) Expansion of the biocyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res* **33**(19):6083–6089, doi:10.1093/nar/gki892.

KAUFFMAN S (1993) The origins of order: Self-organization and selection in evolution. Oxford University Press, New York.

KEPLER TB, CROSBY L, and MORGAN KT (2002) Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biol* **3**(7):RESEARCH0037.

KERRIEN S, ET AL. (2007) Intact–open source resource for molecular interaction data. *Nucleic Acids Res* **35**(Database issue):D561–D565, doi:10.1093/nar/gkl958.

KIEHL TR, MATTHEYSES RM, and SIMMONS MK (2004) Hybrid simulation of cellular behavior. *Bioinformatics* **20**(3):316–322, doi:10.1093/bioinformatics/btg409.

KITANO H (2002) Computational systems biology. *Nature* **420**(6912):206–10, doi:10.1038/nature01254.

KITANO H (2003) A graphical notation for biochemical networks. *biosilico* **1**:169–176.

KITANO H, FUNAHASHI A, MATSUOKA Y, and ODA K (2005) Using process diagrams for the graphical representation of biological networks. *Nat Biotechnol* **23**(8):961–966, doi:10.1038/nbt1111.

KLIPP E, HERWIG R, KOWALD A, WIERLING C, and LEHRACH H (2005) Systems Biology in Practice. Concepts, Implementation and Application. WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

KLOSE J (1975) Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik* **26**(3):231–43.

KLOSE J, ET AL. (2002) Genetic analysis of the mouse brain proteome. *Nat Genet* **30**(4):385–93, doi:10.1038/ng861.

KOHN KW (1999) Molecular interaction map of the mammalian cell cycle control and dna repair systems. *Mol Biol Cell* **10**(8):2703–2734.

LAI EC (2004) Notch signaling: control of cell communication and cell fate. *Development* **131**(5):965–73, doi:10.1242/dev.01074.

LANDER ES, ET AL. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**(6822):860–921.

LEE DY, SAHA R, YUSUFI FNK, PARK W, and KARIMI IA (2008) Web-based applications for building, managing and analysing kinetic models of biological systems. *Brief Bioinform* doi:10.1093/bib/bbn039.

LEE E, SALIC A, KRÜGER R, HEINRICH R, and KIRSCHNER MW (2003) The roles of APC and Axin derived from experimental and theoretical analysis of the Wnt pathway. *PLoS Biol* **1**(1):E10, doi:10.1371/journal.pbio.0000010.

LEHRACH H, DRAMANAC R, HOHEISEL J, LARIN Z, LENNON G, MONACO MP, NIZETIC D, ZEHETNER G, and POUSTKA A (1990) Genome Analysis Volume 1: Genetic and Physical Mapping, chapter Hybridization fingerprinting in genome mapping and sequencing, pp. 39–81. Cold Spring Harbor Laboratory Press, MA.

LENNON GG and LEHRACH H (1991) Hybridization analyses of arrayed cDNA libraries. *Trends Genet* **7**(10):314–7.

LEWIS J (2003) Autoinhibition with transcriptional delay: a simple mechanism for the zebrafish somitogenesis oscillator. *Curr Biol* **13**(16):1398–408.

LIPSHUTZ RJ, FODOR SP, GINGERAS TR, and LOCKHART DJ (1999) High density synthetic oligonucleotide arrays. *Nat Genet* **21**(1 Suppl):20–4, doi:10.1038/4447.

LLOYD CM, HALSTEAD MDB, and NIELSEN PF (2004) CellML: its future, present and past. *Prog Biophys Mol Biol* **85**(2-3):433–50, doi:10.1016/j.pbiomolbio.2004.01.004.

LOCKHART DJ, DONG H, BYRNE MC, FOLLETTIE MT, GALLO MV, CHEE MS, MITTMANN M, WANG C, KOBAYASHI M, HORTON H, and BROWN EL (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* **14**(13):1675–80, doi:10.1038/nbt1296-1675.

LOGAN CY and NUSSE R (2004) The Wnt signaling pathway in development and disease. *Annu Rev Cell Dev Biol* **20**:781–810, doi:10.1146/annurev.cellbio.20.010403.113126.

MARTIN GR (1998) The roles of fgfs in the early development of vertebrate limbs. *Genes Dev* **12**(11):1571–1586.

MATYS V, ET AL. (2006) Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34**(Database issue):D108–D110, doi:10.1093/nar/gkj143.

MENDES P (1993) Gepasi: a software package for modelling the dynamics, steady states and control of biochemical and other systems. *Comput Appl Biosci* **9**(5):563–71, 0266-7061 Journal Article.

MENDES P (1997) Biochemistry by numbers: simulation of biochemical pathways with gepasi 3. *Trends Biochem Sci* **22**(9):361–3, 0968-0004 Journal Article.

MENDES P and KELL D (1998) Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics* **14**(10):869–83, 1367-4803 Journal Article.

MENDES P, SHA W, and YE K (2003) Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics* **19 Suppl 2**:ii122–ii129.

MICHAELIS L and MENTEN M (1913) Die kinetic der invertinwirkung. *Biochemische Zeitschrift* **49**:334–369.

MOHAMMADI M, DIKIC I, SOROKIN A, BURGESS WH, JAYE M, and SCHLESSINGER J (1996) Identification of six novel autophosphorylation sites on fibroblast growth factor receptor 1 and elucidation of their importance in receptor activation and signal transduction. *Mol Cell Biol* **16**(3):977–989.

MOLES CG, MENDES P, and BANGA JR (2003) Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res* **13**(11):2467–2474, doi:10.1101/gr.1262503.

MOODIE S, SOROKIN A, GORYANIN I, and GHAZAL P (2006) A graphical notation to describe the logical interactions of biological pathways. *J Integr Bioinfo* **3**:36.

MOON RT, KOHN AD, FERRARI GVD, and KAYKAS A (2004) WNT and beta-catenin signalling: diseases and therapies. *Nat Rev Genet* **5**(9):691–701, doi:10.1038/nrg1427.

NAITO AT, AKAZAWA H, TAKANO H, MINAMINO T, NAGAI T, ABURATANI H, and KOMURO I (2005) Phosphatidylinositol 3-kinase-akt pathway plays a critical role in early cardiomyogenesis by regulating canonical wnt signaling. *Circ Res* **97**(2):144–151, doi:10.1161/01.RES.0000175241.92285.f8.

NGUYEN C, ROCHA D, GRANJEAUD S, BALDIT M, BERNARD K, NAQUET P, and JORDAN BR (1995) Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones. *Genomics* **29**(1):207–16.

NOBLE D (2002a) Modeling the heart–from genes to cells to the whole organ. *Science* **295**(5560):1678–1682, doi:10.1126/science.1069881.

NOBLE D (2002b) The rise of computational biology. *Nat Rev Mol Cell Biol* **3**(6):459–463, doi:10.1038/nrm810.

NOBLE D (2006) The music of life. Oxford University Press, New York.

NOVÁK B, TÓTH A, CSIKÁSZ-NAGY A, GYÖRFFY B, TYSON JJ, and NASMYTH K (1999) Finishing the cell cycle. *J Theor Biol* **199**(2):223–233, doi:10.1006/jtbi.1999.0956.

NOVÈRE NL, BORNSTEIN B, BROICHER A, COURTOT M, DONIZELLI M, DHARURI H, LI L, SAURO H, SCHILSTRA M, SHAPIRO B, SNOEP JL, and HUCKA M (2006) Biomodels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res* **34**(Database issue):D689–D691, doi:10.1093/nar/gkj092.

OLIVIER BG and SNOEP JL (2004) Web-based kinetic modelling using JWS Online. *Bioinformatics* **20**(13):2143–4, doi:10.1093/bioinformatics/bth200.

ONG SH, GUY GR, HADARI YR, LAKS S, GOTOH N, SCHLESSINGER J, and LAX I (2000) Frs2 proteins recruit intracellular signaling pathways by binding to diverse targets on fibroblast growth factor and nerve growth factor receptors. *Mol Cell Biol* **20**(3):979–989.

ONG SH, HADARI YR, GOTOH N, GUY GR, SCHLESSINGER J, and LAX I (2001) Stimulation of phosphatidylinositol 3-kinase by fibroblast growth factor receptors is mediated by coordinated recruitment of multiple docking proteins. *Proc Natl Acad Sci U S A* **98**(11):6074–6079, doi:10.1073/pnas.111114298.

ORNITZ DM and ITOH N (2001) Fibroblast growth factors. *Genome Biol* **2**(3):REVIEWS3005.

PACOLD ME, SUIRE S, PERISIC O, LARA-GONZALEZ S, DAVIS CT, WALKER EH, HAWKINS PT, STEPHENS L, ECCLESTON JF, and WILLIAMS RL (2000) Crystal structure and functional analysis of ras binding to its effector phosphoinositide 3-kinase gamma. *Cell* **103**(6):931–943.

PALMEIRIM I, HENRIQUE D, ISH-HOROWICZ D, and POURQUIÉ O (1997) Avian hairy gene expression identifies a molecular clock linked to vertebrate segmentation and somitogenesis. *Cell* **91**(5):639–48.

PARKINSON H, ET AL. (2007) Arrayexpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* **35**(Database issue):D747–D750, doi:10.1093/nar/gkl995.

PAWELETZ CP, CHARBONEAU L, BICHSEL VE, SIMONE NL, CHEN T, GILLESPIE JW, EMMERT-BUCK MR, ROTH MJ, III EFP, and LIOTTA LA (2001) Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene* **20**(16):1981–1989, doi:10.1038/sj.onc.1204265.

PETZOLD L (1983) Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations. *siam j sci stat comput* **4**:136–148.

PFEIFFER T, SANCHEZ-VALDENEBRO I, NUNO J, MONTERO F, and SCHUSTER S (1999) Metatool: For studying metabolic networks. *Bioinformatics* **15**:251–257.

PIRSON I, FORTEMAISON N, JACOBS C, DREMIER S, DUMONT JE, and MAENHAUT C (2000) The visual display of regulatory information and networks. *Trends Cell Biol* **10**(10):404–408.

POURQUIÉ O and TAM PP (2001) A nomenclature for prospective somites and phases of cyclic gene expression in the presomitic mesoderm. *Dev Cell* **1**(5):619–620.

POUSTKA A, POHL T, BARLOW DP, ZEHETNER G, CRAIG A, MICHIELS F, EHRICH E, FRISCHAUF AM, and LEHRACH H (1986) Molecular approaches to mammalian genetics. *Cold Spring Harb Symp Quant Biol* **51 Pt 1**:131–9.

POWELL M (1970) A hybrid method for nonlinear equations. In P Rabinowitz (editor), Numerical Methods for Nonlinear Algebraic Equations, pp. 87–114, Gorden and Breach.

RASER JM and O'SHEA EK (2004) Control of stochasticity in eukaryotic gene expression. *Science* **304**(5678):1811–1814, doi:10.1126/science.1098641.

REBAY I, FLEMING RJ, FEHON RG, CHERBAS L, CHERBAS P, and ARTAVANIS-TSAKONAS S (1991) Specific EGF repeats of Notch mediate interactions with Delta and Serrate: implications for Notch as a multifunctional receptor. *Cell* **67**(4):687–99.

REDER C (1988) Metabolic control theory: a structural approach. *J Theor Biol* **135**(2):175–201.

REYA T and CLEVERS H (2005) Wnt signalling in stem cells and cancer. *Nature* **434**(7035):843–850, doi:10.1038/nature03319.

RICHARDSON MK, ALLEN SP, WRIGHT GM, RAYNAUD A, and HANKEN J (1998) Somite number and vertebrate evolution. *Development* **125**(2):151–160.

RODRIGUEZ-VICIANA P, WARNE PH, DHAND R, VANHAESEBROECK B, GOUT I, FRY MJ, WATERFIELD MD, and DOWNWARD J (1994) Phosphatidylinositol-3-oh kinase as a direct target of ras. *Nature* **370**(6490):527–532, doi:10.1038/370527a0.

SALGADO H, GAMA-CASTRO S, PERALTA-GIL M, DÍAZ-PEREDO E, SÁNCHEZ-SOLANO F, SANTOS-ZAVALETA A, MARTÍNEZ-FLORES I, JIMÉNEZ-JACINTO V, BONAVIDES-MARTÍNEZ C, SEGURA-SALAZAR J, MARTÍNEZ-ANTONIO A, and COLLADO-VIDES J

(2006) Regulondb (version 5.0): Escherichia coli k-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res* **34**(Database issue):D394–D397, doi:10.1093/nar/gkj156.

Salin H, Vujasinovic T, Mazurie A, Maitrejean S, Menini C, Mallet J, and Dumas S (2002) A novel sensitive microarray approach for differential screening using probes labelled with two different radioelements. *Nucleic Acids Res* **30**(4):e17.

Sasai Y, Kageyama R, Tagawa Y, Shigemoto R, and Nakanishi S (1992) Two mammalian helix-loop-helix factors structurally related to drosophila hairy and enhancer of split. *Genes Dev* **6**(12B):2620–2634.

Saupe D (1988) The Science of Fractal Images, chapter Chapter 2: Algorithms for random fractals, pp. 71–136. Springer Verlag.

Schacherer F, Choi C, Götze U, Krull M, Pistor S, and Wingender E (2001) The transpath signal transduction database: a knowledge base on signal transduction networks. *Bioinformatics* **17**(11):1053–1057.

Schena M, Shalon D, Davis RW, and Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**(5235):467–70.

Schena M, Shalon D, Heller R, Chai A, Brown PO, and Davis RW (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci U S A* **93**(20):10614–9.

Schilstra MJ and Bolouri H (2002) The logic of gene regulation. In ICSB ’02: Proceedings of the 3rd International Conference on Systems Biology, pp. 197–198, Karolinska Institute, Stockholm, Sweden.

Schilstra MJ and Nehaniv CL (2008) Bio-logic: gene expression and the laws of combinatorial logic. *Artif Life* **14**(1):121–133, doi:10.1162/artl.2008.14.1.121.

Schlessinger J (2000) Cell signaling by receptor tyrosine kinases. *Cell* **103**(2):211–225.

Schlessinger J, Plotnikov AN, Ibrahimi OA, Eliseenkova AV, Yeh BK, Yayon A, Linhardt RJ, and Mohammadi M (2000) Crystal structure of a ternary fgf-fgfr-heparin complex reveals a dual role for heparin in fgfr binding and dimerization. *Mol Cell* **6**(3):743–750.

SCHOMBURG I, CHANG A, EBELING C, GREMSE M, HELDT C, HUHN G, and SCHOMBURG D (2004) Brenda, the enzyme database: updates and major new developments. *Nucleic Acids Res* **32**(Database issue):D431–D433, doi:10.1093/nar/gkh081.

SCHUCHHARDT J, BEULE D, MALIK A, WOLSKI E, EICKHOFF H, LEHRACH H, and HERZEL H (2000) Normalization strategies for cDNA microarrays. *Nucleic Acids Res* **28**(10):E47.

SCHUSTER S and SCHUSTER R (1991) Detecting strictly detailed balanced subnetworks in open chemical reaction networks. *J Math Chem* **6**:17–40.

SCHWANN T and SCHLEIDEN MJ (1839) Mikroskopische Untersuchungen über die Übereinstimmung in der Struktur und dem Wachstum der Tiere und Pflanzen. Berlin.

SCHWANN T and SCHLEIDEN MJ (1847) Microscopic Investigations on the Accordance in the Structure and Growth of Plants and Animals. Sydenham Society.

SLEPCHENKO BM, SCHAFF JC, MACARA I, and LOEW LM (2003) Quantitative cell biology with the virtual cell. *Trends Cell Biol* **13**(11):570–6, 0962-8924 Journal Article.

SPELLMAN PT, SHERLOCK G, ZHANG MQ, IYER VR, ANDERS K, EISEN MB, BROWN PO, BOTSTEIN D, and FUTCHER B (1998) Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* **9**(12):3273–97, 1059-1524 Journal Article.

STEINFATH M, WRUCK W, SEIDEL H, LEHRACH H, RADELOF U, and O'BRIEN J (2001) Automated image analysis for array hybridization experiments. *Bioinformatics* **17**(7):634–41.

STICKNEY HL, BARRESI MJ, and DEVOTO SH (2000) Somite development in zebrafish. *Dev Dyn* **219**(3):287–303, doi:3.0.CO;2-A.

STOFFERS HJ, SONNHAMMER EL, BLOMMESTIJN GJ, RAAT NJ, and WESTERHOFF HV (1992) Metasim: object-oriented modelling of cell regulation. *Comput Appl Biosci* **8**(5):443–9, 0266-7061 Journal Article.

TAKAHASHI K, ISHIKAWA N, SADAMOTO Y, SASAMOTO H, OHTA S, SHIOZAWA A, MIYOSHI F, NAITO Y, NAKAYAMA Y, and TOMITA M (2003) E-cell 2: Multi-platform e-cell simulation system. *Bioinformatics* **19**(13):1727–1729.

TAM PP (1981) The control of somitogenesis in mouse embryos. *J Embryol Exp Morphol* **65 Suppl**:103–128.

TANG M (2008) The mean and noise of stochastic gene transcription. *J Theor Biol* **253**(2):271–280, doi:10.1016/j.jtbi.2008.03.023.

TEUSINK B, PASSARGE J, REIJENGA CA, ESGALHADO E, VAN DER WEIJDEN CC, SCHEPPER M, WALSH MC, BAKKER BM, VAN DAM K, WESTERHOFF HV, and SNOEP JL (2000) Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? testing biochemistry. *Eur J Biochem* **267**(17):5313–5329.

THISSE B and THISSE C (2005) Functions and regulations of fibroblast growth factor signaling during embryonic development. *Dev Biol* **287**(2):390–402, doi:10.1016/j.ydbio.2005.09.011.

TIANA G, KRISHNA S, PIGOLOTTI S, JENSEN MH, and SNEPPEN K (2007) Oscillations and temporal signalling in cells. *Phys Biol* **4**(2):R1–17, doi:10.1088/1478-3975/4/2/R01.

TOLWINSKI NS and WIESCHAUS E (2004) Rethinking wnt signaling. *Trends Genet* **20**(4):177–181, doi:10.1016/j.tig.2004.02.003.

TOMITA M, HASHIMOTO K, TAKAHASHI K, SHIMIZU TS, MATSUZAKI Y, MIYOSHI F, SAITO K, TANIDA S, YUGI K, VENTER JC, and HUTCHISON R C A (1999) E-cell: software environment for whole-cell simulation. *Bioinformatics* **15**(1):72–84, 1367-4803 Journal Article.

TYSON JJ, NOVAK B, ODELL GM, CHEN K, and THRON CD (1996) Chemical kinetic theory: understanding cell-cycle regulation. *Trends Biochem Sci* **21**(3):89–96.

ULLRICH A and SCHLESSINGER J (1990) Signal transduction by receptors with tyrosine kinase activity. *Cell* **61**(2):203–212.

VASTRIK I, ET AL. (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* **8**(3):R39, doi:10.1186/gb-2007-8-3-r39.

VENTER JC, ET AL. (2001) The sequence of the human genome. *Science* **291**(5507):1304–51, doi:10.1126/science.1058040.

VON MERING C, KRAUSE R, SNEL B, CORNELL M, OLIVER SG, FIELDS S, and BORK P (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**(6887):399–403, doi:10.1038/nature750.

WATERSTON RH, ET AL. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**(6915):520–62.

WELCH BL (1947) The generalization of student's problem when several different population variances are involved. *Biometrika* **34**:28–35.

WHITFIELD ML, SHERLOCK G, SALDANHA AJ, MURRAY JI, BALL CA, ALEXANDER KE, MATESE JC, PEROU CM, HURT MM, BROWN PO, and BOTSTEIN D (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell* **13**(6):1977–2000, doi:10.1091/mbc.02-02-0030.

WIERLING C (2006) Pybios - ein modellierungs- und simulationssystem für komplexe biologische prozesse. In K Kremer and V Macho (editors), In Forschung und wissenschaftliches Rechnen. Beiträge zum Heinz-Billing Preis 2005, volume 69, pp. 53–71, Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen (GWDG).

WIERLING C, HERWIG R, and LEHRACH H (2007) Resources, standards and tools for systems biology. *Brief Funct Genomic Proteomic* doi:10.1093/bfgp/elm027.

WIERLING CK, STEINFATH M, ELGE T, SCHULZE-KREMER S, AANSTAD P, CLARK M, LEHRACH H, and HERWIG R (2002) Simulation of dna array hybridization experiments and evaluation of critical parameters during subsequent image and data analysis. *BMC Bioinformatics* **3**(1):29, 1471-2105 Journal article.

WILKINSON DJ (2006) Stochastic Modelling for Systems Biology. Chapman and Hall/CRC.

WINGENDER E, CHEN X, HEHL R, KARAS H, LIEBICH I, MATYS V, MEINHARDT T, PRÜSS M, REUTER I, and SCHACHERER F (2000) Transfac: an integrated system for gene expression regulation. *Nucleic Acids Res* **28**(1):316–319.

WITTIG U, GOLEBIEWSKI M, KANIA R, KREBS O, MIR S, WEIDEMANN S AAND ANSTEIN, SARIC J, and ROJAS I (2006) Sabio-rk: Integration and curation of reaction kinetics data. In In proceedings of the 3rd International workshop on Data Integration in the Life Sciences 2006 (DILS'06). Hinxton, UK. Lecture Notes in Bioinformatics, volume 4075, pp. 94–103.

WODICKA L, DONG H, MITTMANN M, HO MH, and LOCKHART DJ (1997) Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nat Biotechnol* **15**(13):1359–67, doi:10.1038/nbt1297-1359.

WONG PC, ZHENG H, CHEN H, BECHER MW, SIRINATHSINGHJI DJ, TRUMBAUER ME, CHEN HY, PRICE DL, DER PLOEG LHV, and SISODIA SS (1997) Presenilin 1 is required for notch1 and dii1 expression in the paraxial mesoderm. *Nature* **387**(6630):288–292, doi:10.1038/387288a0.

XENARIOS I, RICE DW, SALWINSKI L, BARON MK, MARCOTTE EM, and EISENBERG D (2000) Dip: the database of interacting proteins. *Nucleic Acids Res* **28**(1):289–291.

YAGIL G and YAGIL E (1971) On the relation between effector concentration and the rate of induced enzyme synthesis. *Biophys J* **11**(1):11–27.

YAYON A, KLAGSBRUN M, ESKO JD, LEDER P, and ORNITZ DM (1991) Cell surface, heparin-like molecules are required for binding of basic fibroblast growth factor to its high affinity receptor. *Cell* **64**(4):841–848.

ZEISER S, RIVERA O, KUTTLER C, HENSE B, LASSER R, and WINKLER G (2008) Oscillations of hes7 caused by negative autoregulation and ubiquitination. *Comput Biol Chem* **32**(1):47–51, doi:10.1016/j.compbiolchem.2007.09.004.

ZHAO F, XUAN Z, LIU L, and ZHANG MQ (2005) Tred: a transcriptional regulatory element database and a platform for in silico gene regulation studies. *Nucleic Acids Res* **33**(Database issue):D103–D107, doi:10.1093/nar/gki004.

Abbreviations

APC	adenomatous polyposis coli; a scaffold protein
bHLH	basic helix-loop-helix; specific DNA-binding motif
CSL	CBF1/Su(H)/LAG1 (CSL) family of transcription factors
CV	coefficient of variation; CV = standard deviation / mean
DAE	differential algebraic equation
DAG	diacylglycerol
DSH	dishevelled protein in mouse
DVL	dishevelled protein in human
EGF	epidermal growth factor
FZ	frizzled; seven transmembrane receptor
GAP	GTPase activating protein; GAPs stimulate the GTPase activity of activated G proteins
GEF	guanine nucleotide exchange factor; GEFs activate G proteins by promoting the exchange of GDP by GTP
GSK3β	glycogen synthase kinase-3β
HTTP	hypertext transfer protocol
IP_3	inositol-1,4,5-trisphosphate
Lef1	lymphoid enhanced-binding factor 1
LPR5/6	low density lipoprotein (LDL) receptor-related proteins 5 and 6; single-pass transmembrane protein
N_{ICD}	Notch intracellular domain

ODE	ordinary differential equation
PCR	polymerase chain reaction
pers. comm.	personal communication
PH	pleckstrin homology domain; some PH domains of intracellular signaling molecules can bind to PI(3,4,5)P_3 produced by PI3-kinase
PI3-kinase	phosphatidylinositol 3-kinase
PI(4,5)P_2	phosphatidylinositol 4,5-bisphosphate
PI(3,4,5)P_3	phosphatidylinositol 3,4,5-trisphosphate
PP2A	protein phosphatase 2A
PSM	Presomitic mesoderm
PTB	phosphotyrosine binding domain
SD	standard deviation
SH2	Src homology 2 domain; protein domain that can bind to phosphorylated tyrosine residues
SH3	Src homology 3 domain; protein domain that can bind to proline-rich motifs in intracellular proteins
Sos	son of sevenless
Tcf	transcription factor 1

A Appendix

A.1 Concepts, Tools, and Methods used for the setup of the computational simulation platforms

This section gives some background information on the concepts, tools, and methods that used for the implementation of the computational modeling and simulation platforms.

A.1.1 Object-oriented programming

The paradigm of object-oriented programming (OOP) is the representation of complex features by computational *objects* that provide the significant data and functionalities of their counterpart in real world, where object *attributes* refer to data and object *methods* refer to functionalities of the real object. Objects with identical attributes and methods, but differing in attribute values are subsumed into *classes*. Thus, classes describe attributes and methods of a group of objects. An object that belongs to a certain class and refers to a specific entity of the real world is also called an *instance*. Thus, the terms object and instance are synonymes. Objects can also refer other objects via their attributes; such relations are called *links* or *associations*. Object classes that summarize attributes and methods that are common among other classes, but which do not directly refer to instances of the real world, are called *abstract classes*. A class can also inherit attributes and methods from another class, and the derived class can define further attributes and methods. This is called *inheritance*.

For instance, let us assume we have a class *cell* that has the methods ’grow’ and ’divide’, and the attribute ’volume’. Each time, when external nutrients are available, ’grow’ is called and changes the value of ’volume’ of the respective cell instance, until a critical volume is reached. When this happens, ’divide’ is called and the cell instance is replaced by two daughter-cell instances, with reduced cell-volumes.

Classes, their attributes, methods, and links, as well as their inheritance structure can be represented by diagrams using the notation defined by the unified modeling language (UML).

A.1.2 Python

Python[1] is an interpreted programming language running on different operating systems. Python permits for several coding styles, like structured or procedural programming, functional programming, or object-oriented programming. An important feature of Python is that it is easily extensible by other compiled programming languages like C, C++ or Fortran. Latter became more and more unpopular because of its syntax, but several Fortran libraries especially for mathematical routines are still in use.

A.1.3 Zope Web Application Server

Zope[2] stands for "Z Object Publishing Environment", and it is a web application server primarily written in the Python programming language. It comprises a Web server, that enables the interaction with the user, and an object-oriented database, that is used by PyBioS to store the models and make their objects persistent. Therefore, no explicite file-format (or table structure for a relational database) is required, since the class definitions and object relations already define the required structure. Zope also maps object methods to incoming HTTP requests and thus it provides dynamic HTML representations of the individual objects.

A.1.4 Numerical Solvers for ODEs and DAEs

The PyBioS modeling and simulation platform that is developed in this thesis and is introduced in section 2.1 can automatically generate a mathematical model, described by an ODE system, from a given topology of a biological model and a set of according kinetic laws. Since these ODE systems often possess non-linear kinetics, in nearly all cases they cannot be solved analytically, but often numerically. PyBioS supports deterministic simulations by numerical integration of first order ODE-systems. It offers the use of the solvers LIMEX and LSODA to get the numerical solution of the initial value problem. LSODA (Hindmarsh, 1983; Petzold, 1983) is a solver for ordinary differential equations written in Fortran and it is a variant of the LSODE package. The algorithm used in this solver switches between stiff and non-stiff methods automatically. PyBioS uses the interface to LSODA which is available from SciPy[3]. The solver LIMEX[4] (Deuflhard et al., 1987; Deuflhard and Nowak, 1987) is an extrapolation integrator for the solution of linearly-implicit differential-algebraic systems (DAEs) written in Fortran. It combines an implicit one step method with stepsize extrapolation to permit an adaptive control of stepsize and order.

[1] http://www.python.org/

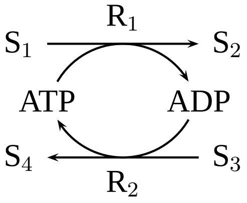
[2] http://www.zope.org/

[3] http://www.scipy.org

[4] ftp://elib.zib.de/pub/elib/codelib/LIMEX4.2A1

A.1.5 Computation of Conservation Relations

For a biochemical reaction the reaction equation describes which molecules are consumed or produced in the reaction and with which molecularities they participate. For a system of reactions this can be described by the stoichiometric matrix. It is a matrix of the stoichiometric coefficients in which each line corresponds to a component and each row corresponds to a reaction, e.g., for the following system of reactions



(A.1)

the stoichiometric matrix $\boldsymbol{N}$ reads

$$\boldsymbol{N} = \begin{array}{c|cc|} & \mathrm{R}_1 & \mathrm{R}_2 \\ \mathrm{S}_1 & -1 & 0 \\ \mathrm{S}_2 & 1 & 0 \\ \mathrm{S}_3 & 0 & -1 \\ \mathrm{S}_4 & 0 & 1 \\ \mathrm{ATP} & -1 & 1 \\ \mathrm{ADP} & 1 & -1 \end{array}. \tag{A.2}$$

Using this notation the system equations (cf. equation 1.9 on page 19) can also be written as

$$\frac{d\boldsymbol{S}}{dt} = \boldsymbol{N}\boldsymbol{v}(\boldsymbol{S}, \boldsymbol{p}), \tag{A.3}$$

where $\boldsymbol{S} = (S_1, S_2, \dots, S_n)^{\mathrm{T}}$ is a vector of the concentrations of the substances, $\boldsymbol{v} = (v_1, v_2, \dots, v_r)^{\mathrm{T}}$ a vector of reaction rates, and $\boldsymbol{p} = (p_1, p_2, \dots, p_m)^{\mathrm{T}}$ a vector of the parameters.

The model in reaction A.1 shows an interesting property of reaction networks that frequently occurs. ATP and ADP are always converted into each other without changing its total amount. Such cycles, called moiety-conserved cycles, arise when groups of atoms, termed moieties, migrate through the network without being synthesized or degraded (Klipp et al., 2005).

The conservation relations described by the moiety-conserved cycles reveal as linear dependencies of the rows of the stoichiometric matrix.

The mathematical derivation of conservation relations—as it is implemented in PyBioS—is done as described by Klipp et al. (2005, pp. 165): A matrix $\boldsymbol{G}$ is considered that fulfills

$$\boldsymbol{G}\boldsymbol{N} = 0. \tag{A.4}$$

Due to Equation (A.3) it follows that

$$\boldsymbol{G}\dot{\boldsymbol{S}} = \boldsymbol{G}\boldsymbol{N}\boldsymbol{v} = 0. \tag{A.5}$$

Integrating this equation leads to the conservation realations

$$\boldsymbol{G}\boldsymbol{S} = \text{const.} \tag{A.6}$$

The conservation matrix $\boldsymbol{G}$ can be calculated from

$$\boldsymbol{N}^{\mathrm{T}}\boldsymbol{G}^{\mathrm{T}} = 0 \tag{A.7}$$

using the block diagonalization algorithm described by (Schuster and Schuster, 1991).

Conservation relations can be used to simplify the system of differential equations $\dot{\boldsymbol{S}} = \boldsymbol{N}\boldsymbol{v}$ that describe the dynamics of the reaction system. This can be done by eliminating linear dependent differential equations and replacing them by appropriate algebraic equations.

The procedure looks as follows (Reder, 1988): Rows of the stoichiometric matrix $\boldsymbol{N}$ and of the concentration vector $\boldsymbol{S}$ have to be reordered in such a way that a set of independent rows is on the top and the dependent rows are at the bottom. Then the matrix $\boldsymbol{N}$ is split into the independent part $\boldsymbol{N}^0$ and the dependent part $\boldsymbol{N}'$, and a link matrix is introduced in the following way

$$\boldsymbol{N} = \begin{pmatrix} \boldsymbol{N}^0 \\ \boldsymbol{N}' \end{pmatrix} = \boldsymbol{L}\boldsymbol{N}^0 = \begin{pmatrix} \boldsymbol{I}_{\mathrm{rank}(\boldsymbol{N})} \\ \boldsymbol{L}' \end{pmatrix} \boldsymbol{N}^0. \tag{A.8}$$

$\boldsymbol{I}_{\mathrm{rank}(\boldsymbol{N})}$ is the identity matrix of size $\mathrm{rank}(\boldsymbol{N})$. The differential equation system may be rewritten accordingly

$$\dot{\boldsymbol{S}} = \begin{pmatrix} \dot{\boldsymbol{S}}_{\mathrm{indep}} \\ \dot{\boldsymbol{S}}_{\mathrm{dep}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_{\mathrm{rank}(\boldsymbol{N})} \\ \boldsymbol{L}' \end{pmatrix} \boldsymbol{N}^0\boldsymbol{v}, \tag{A.9}$$

and the dependent concentrations fulfill

$$\dot{\boldsymbol{S}}_{\mathrm{dep}} = \boldsymbol{L}' \cdot \dot{\boldsymbol{S}}_{\mathrm{indep}} + \text{const.} \tag{A.10}$$

This relation is fulfilled during the entire time course. Thus we may replace the original system by a reduced differential equation system

$$\dot{\boldsymbol{S}}_{\mathrm{indep}} = \boldsymbol{N}^0\boldsymbol{v} \tag{A.11}$$

supplemented with the set of algebraic equations (Eq. (A.10)).

A.2 Modeling of Somitogenesis

A.2.1 Kintics Used Within the Somitogenesis Model

Complex Dissociation Complex dissociation is described by a kinetic law of a reversible reaction.

$$\mathrm{AB} \rightleftharpoons \mathrm{A + B} \tag{A.12}$$

The corresponding rate law is

$$v = k_{\mathrm{off}}[AB] - k_{\mathrm{off}}/k_{\mathrm{D}}[A][B], \tag{A.13}$$

where $k_{\mathrm{D}} = k_{\mathrm{off}}/k_{\mathrm{on}}$ is the dissociation constant, and k_{on} and k_{off} are the association and dissociation rate constants, respectively.

Complex Association Complex association is described by a kinetic law of a reversible reaction.

$$\mathrm{A + B} \rightleftharpoons \mathrm{AB} \tag{A.14}$$

The corresponding rate law is

$$v = k_{\mathrm{on}}[A][B] - k_{\mathrm{on}} \cdot k_{\mathrm{D}}[AB], \tag{A.15}$$

where $k_{\mathrm{D}} = k_{\mathrm{off}}/k_{\mathrm{on}}$ is the dissociation constant, and k_{on} and k_{off} are the association and dissociation rate constants, respectively.

Degradation Reactions Degradation of proteins, mRNAs or complexes are described either by a first-order reaction or by a Michaelis-Menten reaction.

$$\mathrm{A} \longrightarrow \tag{A.16}$$

The corresponding rate law of a first-order reaction is

$$v = k[A], \tag{A.17}$$

where k is the first order rate constant or

$$v = \frac{V_{max}\,[S]}{[S] + K_m}, \tag{A.18}$$

where V_{max} is the maximal rate of the reaction and K_m is the substrate concentration for which the reaction rate is half maximal.

Synthesis Reactions Synthesis reactions of proteins are either described by a zero-order reaction, a first-order reaction

For a zero-order reaction the reaction rate is

$$v = k, \tag{A.19}$$

where k is the reaction rate coefficient.

Gene expression processes are described as follows.

Single activator:

$$v_i = V \cdot \frac{[A]^n}{[A]^n + K_a^n} + b \tag{A.20}$$

One activator and one inhibitor:

$$v_i = V_i \cdot \prod_j \left(\frac{K_{i_j}^{n_j}}{I_j^{n_j} + K_{i_j}^{n_j}} \right) \times \prod_k \left(\frac{A_k^{n_k}}{A_k^{n_k} + K_{a_k}^{n_k}} \right) + b \tag{A.21}$$

Two activators and one inhibitor:

$$v_i = V \cdot \prod_j \left(\frac{K_{i_j}^{n_j}}{I_j^{n_j} + K_{i_j}^{n_j}} \right) \times \prod_k \left(\frac{A_k^{n_k}}{A_k^{n_k} + K_{a_k}^{n_k}} \right) \left(\frac{A_k^{n_k}}{A_k^{n_k} + K_{a_k}^{n_k}} \right) \tag{A.22}$$

A.3 Modeling of DNA Arrays

A.3.1 cDNA Array Data Used for Modeling

In section 2.3 I present a study concerning the evaluation of critical parameters occuring in DNA array hybridization experiments. I simulated hybridized filter images according to different sources of error and used them for subsequent analysis. In DNA array experiments errors might arose from variations of the spot positions caused by different experimental artifacts or by different sources of background noise. To use a realistic distribution of signals as input data for the simulations, intensity values and their respective grid positions were taken from experiments with macroarrays. The origin of the experimental data is described in the following.

A detailed description of the cDNA clone array design, mRNA labeling, hybridization and data capture is given in Herwig et al. (2001). PCR products of 14 208 zebrafish cDNA clones of a representative library from gastrula stage embryos (Clark et al., 2001) and 2 304 copies of an *Arabidopsis thaliana* cDNA clone were spotted on nylon filter membranes. Clones were spotted in a rectangular grid of blocks with 25 spots (5×5) per block by the use of a gadget with 16×24 pins corresponding to a 384-well microtiter plate. Figure 2.16B on page 68 illustrates the filter design. Due to the experimental procedure a filter is divided into six fields of 384 blocks each. For the 5×5 spotting pattern each block comprises 25 spots. The zebrafish target derived from mRNA of gastrula stage embryos (6 hours post fertilization) was hybridized to six filter replicates which were spotted with the same set of clones. To improve reproducibility, each clone was spotted in duplicate per block. The spot intensities of the hybridized filters were analyzed as described in Herwig et al. (2001). For each spotted cDNA clone mean signal intensities were calculated from the six filter replicates and used as input data for the simulations in section 2.3.

A.3.2 Data acquisition in DNA array experiments

Image analysis is the first bioinformatics module in the data analysis pipeline of DNA array experiments. In this step each probe spot of the scanned DNA array image is assigned a numerical value that represents the signal intensity. Essential for this is the correct identification of each spot center and a correct quantitation of the pixel neighborhood around the identified center of each spot. Since the signal intensities determined during image analysis are the input data to any further pre-processing steps and fold-change analysis or clustering analysis, the quality of image analysis is essential for any results that can be gaind by subsequent procedures. In section 2.3 simulated images that represent different experimental errors are used to study how the degree of automation of the image analysis affects the quality of image analysis. Image analysis methods can be grouped into three classes: manual, semiautomated, and automated methods. Manual methods strongly rely on supervision of the user by requiring an inital guess on the spot positions, e.g., the user has to adjust an ideal grid manually on the screen. Semiautomated methods require less interaction, but still need some prior information, e.g., the definition of the spotted area. Automated methods try to find the spot grid without any user interaction. For the mentioned study, three programs were chosen to represent each of the three classes:

Visual Grid This program is a commercial product of the company GPC Biotech AG[5]. The program provides the functionality to individually define the grid, sub-grid, and each spot position by the user. Since the whole grid has to be adapted manually, its degree of automation can be classified as ’manual’.

Aida This program is a commercial product of the company Raytest[6]. It requires only a limited interaction by the user for the grid positioning; a fine-tuning of the spot positions is performed automatically. Thus, it can be classified as ’semiautomated’.

[5] `http://www.gpc-ag.com`

[6] `http://www.raytest.de`

Filter-Analysis tool FA The third program is the filter analysis tool FA, developed by Steinfath et al. (2001) at the Max Planck Institute for Molecular Genetics. It uses an algorithm for the grid detection that requires no interaction by the user. The program automatically identifies the global borders of the rectangular grid. In a step-down procedure it detects subgrids, and finally performs also a fine-tuning for each spot position. Thus, it can be classified as 'automated'.

A.3.3 Mathematical Description of a Crater Spot

A crater spot can be described by the following function:

$$f(x) = \frac{1}{2\pi\sigma_1^2} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma_1}\right)^2} - \frac{1}{2\pi\sigma_2^2} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma_2}\right)^2} \tag{A.23}$$

$$g(x) = a \cdot e^{bh^2(x)} \tag{A.24}$$

$$g'(x) = a \cdot e^{bh^2(x)} \cdot 2bh(x) \cdot h'(x) \tag{A.25}$$

$$a1 = \frac{1}{2\pi\sigma_1^2}; \qquad b1 = -\frac{1}{2\sigma_1^2} \tag{A.26}$$

$$a2 = \frac{1}{2\pi\sigma_2^2}; \qquad b1 = -\frac{1}{2\sigma_2^2} \tag{A.27}$$

$$h^2(x) = (x-\mu)^2; h = x - \mu; h'(x) = 1 \tag{A.28}$$

The derivation of f is given by

$$f'(x) = g_1'(x) - g_2'(x) \tag{A.29}$$

$$= \frac{1}{2\pi\sigma_1^2} \cdot e^{\frac{(x-\mu)^2}{2\sigma_1^2}} \cdot \left(-\frac{1}{\sigma_1^2}\right)(x-\mu) - \tag{A.30}$$

$$\left(\frac{S}{2\pi\sigma_2^2} \cdot e^{-\frac{(x-\mu)^2}{2\sigma_2^2}} \cdot \left(-\frac{1}{\sigma_2^2}\right)(x-\mu)\right) \tag{A.31}$$

$$= -\frac{x-\mu}{2\pi\sigma_1^4} \cdot e^{-\frac{(x-\mu)^2}{2\sigma_1^2}} + S\frac{x-\mu}{2\pi\sigma_2^4} \cdot e^{-\frac{(x-\mu)^2}{2\sigma_2^2}} \tag{A.32}$$

$$f'(x) = 0 \tag{A.33}$$

$$\frac{x-\mu}{2\pi\sigma_1^4}\cdot e^{-\frac{(x-\mu)^2}{2\sigma_1^2}} = S\frac{x-\mu}{2\pi\sigma_2^4}\cdot e^{-\frac{(x-\mu)^2}{2\sigma_2^2}} \quad ; \quad x \neq \mu \tag{A.34}$$

$$\frac{1}{\sigma_1^4}\cdot e^{-\frac{(x-\mu)^2}{2\sigma_1^2}} = \frac{S}{\sigma_2^4}\dot{e}^{-\frac{(x-\mu)^2}{2\sigma_2^2}} \tag{A.35}$$

$$\ln\frac{\sigma_2^4}{S\sigma_1^4} = \ln\frac{e^{-\frac{(x-\mu)^2}{2\sigma_2^2}}}{e^{-\frac{(x-\mu)^2}{2\sigma_1^2}}} = -\frac{(x-\mu)^2}{2\sigma_2^2} + \frac{(x-\mu)^2}{2\sigma_1^2} \tag{A.36}$$

$$2\ln\frac{\sigma_2^4}{S\sigma_1^4} = \frac{x^2-2\mu x+\mu^2}{\sigma_1^2} - \frac{x^2-2\mu x+\mu^2}{\sigma_2^2} \tag{A.37}$$

$$2\sigma_1^2\sigma_2^2\ln\frac{\sigma_2^4}{S\sigma_1^4} = \sigma_2^2x^2 - 2\sigma_2^2\mu x + \mu^2\sigma_2^2 - \sigma_1^2x^2 + 2\mu\sigma_1^2x - \sigma_1^2\mu^2 \tag{A.38}$$

$$2\frac{\sigma_1^2\sigma_2^2}{\sigma_2^2-\sigma_1^2}\ln\frac{\sigma_2^4}{S\sigma_1^4} = x^2 - 2\mu x + \mu^2 \tag{A.39}$$

$$0 = x^2 - 2\mu x + \mu^2 - 2\frac{\sigma_1^2\sigma_2^2}{2\sigma_2^2-\sigma_1^2}\ln\frac{\sigma_2^4}{S\sigma_1^4} \tag{A.40}$$

$$x_{1/2} = -\frac{p}{2} \pm \sqrt{\left(\frac{p}{2}\right)^2 - q} \tag{A.41}$$

$$x_{1/2} = \mu \pm \sqrt{2\frac{(\sigma_1\sigma_2)^2}{\sigma_2^2-\sigma_1^2}\ln\frac{\sigma_2^4}{S\sigma_1^4}} \tag{A.42}$$

$$x_{1/2} = \mu \pm \left(\sigma_1\sigma_2\sqrt{\frac{2}{\sigma_2^2-\sigma_2^2}\ln\frac{\sigma_2^4}{S\sigma_1^4}}\right) \tag{A.43}$$

For which values is a crater defined: ($\sigma_1 > \sigma_2$; because for these a crater-rim—local max.—is defined)

$$\frac{2}{\sigma_2^2-\sigma_1^2}\ln\frac{\sigma_2^4}{S\sigma_1^4} > 0 \quad ;\sigma_1 > \sigma_2 \tag{A.44}$$

$$\ln\frac{\sigma_2^4}{S\sigma_1^4} < 0 \tag{A.45}$$

$$\frac{\sigma_2^4}{S\sigma_1^4} < 1 \tag{A.46}$$

$$\sigma_2^4 < S\sigma_1^4 \tag{A.47}$$

$$\sigma_2 < \sqrt[4]{S\sigma_1^4} = \sqrt[4]{S}\sigma_1 \quad ;S < 1 \tag{A.48}$$

The crater does not become negative, if

$$\frac{1}{2\pi\sigma_1^2} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma_1}\right)^2} - S\frac{1}{2\pi\sigma_2^2} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma_2}\right)^2} > 0 \quad ; for x = \mu \tag{A.49}$$

$$\frac{1}{\sigma_1^2} > \frac{S}{\sigma_2^2} \tag{A.50}$$

$$\sigma_2 > \sqrt{\sigma_1^2 S} \tag{A.51}$$

A crater is defined, if

$$\sqrt{S}\sigma_1 < \sigma_2 < \sqrt[4]{S}\sigma_1 \tag{A.52}$$

If σ_1 and S are given, σ_2 can be calclated for a given radius r as follows:

$$r = \sigma_1\sigma_2\sqrt{\frac{2}{\sigma_2^2 - \sigma_1^2}\ln\frac{\sigma_2^4}{S\sigma_1^4}} \quad ; \sigma_2 = \sigma_1\sqrt{S} \tag{A.53}$$

$$r = \sigma_1^2\sqrt{S}\sqrt{\frac{2}{\sigma_1^2 S - \sigma_1^2}\ln S} \tag{A.54}$$

$$r = \sqrt{\frac{2S\sigma_1^2}{S-1}\ln S} \tag{A.55}$$

$$r^2 = \frac{2S\sigma_1^2}{S-1}\ln S \tag{A.56}$$

$$\sqrt{\frac{r^2(S-1)}{2\ln(S)S}} = \sigma_1 = \frac{\sigma_2}{\sqrt{S}} \tag{A.57}$$

$$\sigma_2 = \sqrt{\frac{r^2(S-1)}{2\ln S}} \text{and} \sigma_1 = \frac{\sigma_2}{\sqrt{S}} \tag{A.58}$$

Publications

Books

Klipp, E., Herwig, R., Kowald, A., Wierling, C., and Lehrach, H. (2005) Systems Biology in Practice: Concepts, Implementation and Application. Wiley-VCH, Weinheim.

Klipp, E., Liebermeister, W., Wierling, C., Kowald, A., Lehrach, H., and Herwig, R. (2009) Systems Biology - A textbook. Wiley-VCH, Weinheim.

Papers

Nebrich, G., Herrmann, M., Hartl, D., Diedrich, M., Kreitler, T., Wierling, C., Klose, J., Giavalisco, P., Zabel, C., and Mao, L. (2009) PROTEOMER: A Workflow-Optimized Laboratory Information Management System (LIMS) for 2-D-Electrophoresis-centered Proteomics *Proteomics*, **9**(7):1795-1808.

Kamburov, A., Wierling, C., Lehrach, H., and Herwig, R. (2009) ConsensusPathDB–a database for integrating human functional interaction networks. *Nucleic Acids Research*, **37**:Database issue D623-D628.

Wierling, C., Herwig, R., and Lehrach, H. (2007) Resources, Standards and Tools for Systems Biology. *Briefings in Functional Genomics and Proteomics*, **6**:240-251.

Petrov V., Wierling C., Maschke-Dutz E., Herwig R. (2007) Qualitative Analysis of Somitogenesis Models. *Bioautomation*, **8**Suppl.1:95-104.

Hache, H., Wierling, C., Lehrach, H., Herwig, R. (2007) Reconstruction and Validation of Gene Regulatory Networks with Neural Networks. Proceedings of the 2nd Foundations of Systems Biology in Engineering Conference (FOSBE), Stuttgart, pp. 319-324.

Wierling, C. (2006) PyBioS - ein Modellierungs- und Simulationssystem für komplexe biologische Prozesse. In Forschung und wissenschaftliches Rechnen. Beiträge zum Heinz-Billing Preis 2005 (Hrsg. K. Kremer, V. Macho), **69**:53-71. Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen (GWDG).

Wierling, C. and Herwig, R. (2006) Systems biology and drug discovery - A new path in genome research. *Screening - Trends in drug discovery*, **7**(3):39-41.

Wierling, C. and Herwig, R. (2006) Vom Gen zum System. *GIT Labor-Fachzeitschrift*, **50**(5):484-486.

Adjaye, J., Huntriss, J., Herwig, R., BenKahla, A., Brink, T.C., Wierling, C., Hultschig, C., Groth, D., Yaspo, M.-L., Picton, H.M., Lanzendorf, S., Gosden, R.G., and Lehrach, H.(2005) Primary differentiation in the human blastocyst: Comparative molecular portraits of inner cell mass and trophectoderm cells. *Stem Cells*, **23**:1514-1525.

Klipp, E., Liebermeister, W., and Wierling, C. (2004) Inferring dynamic properties of biochemical reaction networks from structural knowledge. *Genome Informatics* **15**(1):125-37.

Wierling, C.K., Steinfath, M., Elge, T., Schulze-Kremer, S., Aanstad, P., Clark, M., Lehrach, H., and Herwig, R. (2002) Simulation of DNA array hybridization experiments and evaluation of critical parameters during subsequent image and data analysis. *BMC Bioinformatics* **3**:29.

Guerasimova, A., Nyarsik, L., Girnus, I., Steinfath, M., Wruck, W., Griffiths, H., Herwig, R., Wierling, C., O'Brien, J., Eickhoff, H., Lehrach, H., and Radelof, U. (2001) New tools for oligonucleotide fingerprinting. *BioTechniques* **31**(3):490–495.

Acknowledgments

I am very grateful to my adviser, Prof. Dr. Hans Lehrach, for giving me the chance to write my PhD thesis in his department and to work on the novel and highly interesting field of systems biology.

My sincere thanks go to Prof. Dr. Volker Erdmann for reviewing this thesis.

At this point, I would like to express my gratitude to Dr. Ralf Herwig who was always ready to give me his support and very helpful and constructive advice on all aspects of my scientific research.

Moreover, I am very grateful to Prof. Dr. Bernhard Herrmann for introducing me to and guiding me through the very interesting field of somitogenesis and all the different regulatory pathways coming along with it.

Furthermore, I am indebted to the Bioinformatics group of the department Prof. Lehrach, namely Marcus Albrecht, Andriani Daskalaki, Felix Dreher, Mario Drungowski, Hendrik Hache, Atanas Kamburov, Elisabeth Maschke-Dutz, Thomas Meinel, Axel Rasche, Anja Thormann, Wasco Wruck, Lukas Chavez Wurm, and Reha Yildirimman as well as the former members Claudia Schepers, Matthias Steinfath, and Thomas Kreitler.

I also thank Michal-Ruth Schweiger, Hendrik Hache, Felix Dreher, Ralf Herwig, Marcus Albrecht, and Atanas Kamburov for proof-reading and commenting on the manuscript of my thesis.

Finally, my special thanks go to my dear family and friends.

This work was supported by the Max Planck Society.