

An Introduction to Systems Biology: Opportunities and Challenges for Physical Scientists in the Post-Genome Era of Biology

Olaf Wolkenhauer^{†*} and Hiroaki Kitano[‡]

[†]Department of Biomolecular Sciences and Department of Electrical Engineering & Electronics
Address: Control Systems Centre, UMIST, Manchester M60 1QD, UK

[‡]Sony Computer Science Laboratories, 3-14-13 Higashi-Gotanda, Shinagawa-Ku, Tokyo 141-0022, Japan

Abstract

Systems theory or systems science has never really managed to achieve widespread and independent status in curricula, departments and journals but instead acts as an umbrella for a number of research activities across the physical and engineering sciences. Now, with revolutionary developments in the life sciences, there is a renewed interest into systems thinking. In this article we survey opportunities and challenges for the application of system theory to genomics – a new area of research also referred to as *systems biology*.

Introduction

With the sequencing of DNA for a number of genomes scientists now have an inventory of genes available to embark on the study of the organisation and control of genetic pathways. This new phase in this biological revolution, the post-genome era, is closely associated with the field of 'functional genomics'. Genomics takes us from the DNA sequence of a gene to the structure of the product for which it codes (usually a protein) to the activity of that protein and its function within a cell, the tissue and ultimately, the organism. A series of articles in Nature [Nature 2000] are recommended for an introduction to the areas of genomics, proteomics and transcriptomics.

With the emergence of genomics, molecular biology currently witnesses a shift of focus from molecular characterisation to the understanding of functional activity. The two central questions biologists investigate are "What are the genes' functional role?" and "How do genes and/or proteins interact?". In the past single genes were studied but with DNA microarray technology we can measure the activity levels of thousands of genes at the same time. It becomes therefore possible to identify interrelationships between groups of genes (with respect to their functional role) and to analyse dynamic interactions among genes ("gene networks"). Similar, proteomics research shows that most proteins interact with several other proteins and it is increasingly appreciated that the function of a protein is appropriately described in the context of its interactions with other proteins. Most of these relationships are dynamic and controlled processes and it is not surprising that there is a renewed interest in the application of systems thinking to biology.

The outline of this article is as follows. We first introduce systems biology in the context of the study of complex systems, reviewing a number of related and relevant areas of research and define complexity in the context of biological systems. Systems biology has a history, and in its early stages involved eminent researchers including Wiener, Kalman, Bertalanffy, Rosen and

* This article is, in parts, based on conclusions reached from the Systems Theory & Genomics research network, funded by the UK Engineering and Physical Sciences Research Council (EPSRC). Author to whom correspondence should be addressed: olaf.wolkenhauer@umist.ac.uk

Mesarovic in the 1960's. We discuss why these attempts vanished from the research agendas and why there is a renewed interest in the post-genome era of the life sciences. This is followed by an example of bacterial gene expression and regulation before outlining current activities from groups around the world. We conclude by listing some of the challenges and hurdles for this (re-)emerging field.

Genomic Cybernetics

The understanding of causality and coping with complexity is not the holy grail of science but part of its very definition. Not surprisingly then, complexity studies have remained as elusive as inconclusive.

Warren Weaver [1948] defined 'disorganised complexity' as a problem in which the number of variables is very large, and any of these variables is best described as a random process. Here we are at the 'molecular level' and the most successful formal methods in representing phenomena at this level derive from statistical considerations. In the context of the cell, at the 'cellular level', matters are complicated by the fact that '*organisation*' becomes an essential feature of the processes under consideration. Weaver referred to problems in which a large number of factors are interrelated into a whole as '*organised complexity*'. The number of variables is too large to be dealt with in the Newtonian realm of physics and mathematical modelling, and the systems are too organised to allow a statistical techniques either. He described organised complexity as the challenge for science in the coming 50 years. His enthusiasm expressed then, is very much how one feels today in the post-genome era of the life-sciences: "It is doubtless true that we are only scratching the surface of the cancer problem, but at least there are now some tools to dig with and there have been located some spots beneath which almost surely there is pay-dirt." [Weaver 1948]

Following Herman Haken's synergetics, chaos-theory, and fractals, the science of self-organised criticality [Bak 1997], non-equilibrium physics, power laws and emergent phenomena renewed the interest in complexity studies over the last decade or so. These studies developed mostly within the areas of physics and mathematics. They are seeking general principle of phenomena that can be observed in a wide range of disciplines. Although frequently motivated by biological examples, complexity studies have failed to have an impact on biology.

Stuart Kauffman's work on genetic networks [Kauffman 1995] paved the way of complexity studies in biology. The work of Brian Goodwin [1994, Sole 2000], Harrison [1993] and Meinhardt [1998] marked a trend to an approach more focussed on specific organisms but continued to investigate cellular processes and morphological development in evolutionary terms. As Harold pointed out in his recent book: "complexity studies is a fresh label for a well known pigeonhole: general systems theory, that was pioneered by Ludwig von Bertalanffy in the thirties." [Harold 2001, p.222]. Systems and control theory on the other hand have their roots in Norbert Wiener's Cybernetics. Systems Biology is therefore an emerging field that continues this research into the post-genome era of the life sciences [Kitano 2001, 2002, Wolkenhauer 2001]. The most significant differences, however, to complexity studies are that systems biology takes a signal- and systems oriented approach, and validates models with experimental data and findings in functional genomics. As we shall see further below, it has more to do with the application of systems and control theory to cellular systems than with the application of physics to biology. Systems biology provides a vital interface between cell biology and biotechnological applications. Before we introduce this area in greater detail in subsequent sections, we note that complexity in the context of biological systems can be defined as

1. A property of an encoding (mathematical model), e.g., its dimensionality, order or number of variables.

2. An attribute of the natural system under consideration, e.g., the number of components, descriptive and organisational levels that ensure its integrity.
3. Our ability to interact with the system, to observe it, i.e., to make measurements and generate experimental data.

On all three accounts, genes, cells, tissue, organs, organisms and populations are individually and as a functional whole a complex system. It is the availability of experimental techniques, modern microscopy, laser tweezers, nanotechnology as well as DNA microarrays, gel technology, and mass spectrometry, which drive this renewed interest in complexity studies and systems biology [Kitano 2001, 2002]. While the technology to generate and manage data races ahead, it becomes apparent that methodological advances in the analysis of data are urgently required if we want to turn the newly available data into information and knowledge. This need for research into new methodologies and the development of novel conceptual frameworks has been neglected in the euphoria about new technology. Problems in the post-genome era of the life sciences will not only be experimental or technical but also conceptual. The interpretation of data, turning information into knowledge is as important for scientific and biotechnological progress as the possibility of generating and managing data.

With the generation of vast amounts of data, computer scientists have been the natural allies of biologists in the management of these data. The growth of bioinformatics parallels the exciting developments in biology. The availability of genome sequence data has led to a shift of focus from molecular characterisation and sequence analysis to an understanding of functional activity and now interactions of genes and proteins in pathways. Gene expression and regulation, to understand the organisation and dynamics of genetic-, signalling- and metabolic pathways is the challenge for the next 50 years. The nature of the experiments and the data thereby generated requires an alliance of the biological and biomedical sciences with physical scientists (engineers, mathematicians and physicists). From the following discussion about the challenges and hurdles, it will become clear why such an alliance is so important.

(Not) A New Kid on the Block

Although, generally considered to be a new area of research, systems biology is not without history and as early as the 1960's the term was used to describe the application of systems and control theory to biology [Wolkenhauer 2001]. At the time, Mihajlo Mesarovic wrote: "In spite of the considerable interest and efforts, the application of systems theory in biology has not quite lived up to expectations. [...] one of the main reasons for the existing lag is that systems theory has not been directly concerned with some of the problems of vital importance in biology." Today, scientists in this field are motivated by the availability of experimental data, including, for example, DNA microarray time series and interdisciplinary collaborations are widely supported. In fact, the importance of interdisciplinary research and close collaborations between biologists and physical scientists is evident in the many multidisciplinary research centres that are build around the world, gently forcing researchers to interact by confining them into purpose-built housing.

Mesarovic further suggested that progress could be made by more direct and stronger interactions of biologists with system scientists: "The real advance in the application of systems theory to biology will come about only when the biologists start *asking questions* which are based on the system-theoretic concepts rather than using these concepts to represent in still another way the phenomena which are already explained in terms of biophysical or biochemical principles. [...] then we will not have the 'application of engineering principles to biological problems' but rather a field of *systems biology* with its own identity and in its own right." Molecular characterisation has lead to very accurate spatial representations of cellular components and biochemical modelling has been the main approach to study cellular

processes. However, the future lies in extending this knowledge to observations at higher organisational levels. There are few examples of a concerted effort to 'translate' biological representations of gene expression and regulation into the language of the system scientist [Kremling 2000 and <http://bioinformatics.org/sbw/>] and all indications are that the field is going to provide the vital interface between basic cell biology, physiology and biotechnological applications such as for example in metabolic engineering.

An important difference is that systems biology has available new technologies to generate data from the genome, transcriptome, proteome, metabolome in addition to existing data from the physiome. The area of systems biology is therefore the most focussed and applied approach to complexity studies, applied to biology, yet. Biologists are generating facts at an unprecedented rate but as Henri Poincaré said in 1913, "Science is built up of facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones is a house." The renewed interest in complexity studies, the motivation for systems biology, is therefore closely linked to the desire to integrate the available information from novel experimental technologies in genomics and proteomics.

Systems Biology is a significant advance from bioinformatics with its focus on molecular characterisation, sequence analysis and data management. Systems theory is not a collection of facts but a way of thinking. While we may never be able to build accurate predictive models of cellular or genetic systems, the modelling process itself will prove valuable to the biologist, helping him to identify which variables to measure and why. The quest for precision is analogous to the quest for certainty and both - precision and certainty are impossible to attain, at present if not in general.

Gene Expression and Regulation: An Example

Each cell of a (multicellular) organism holds the genome with the entire genetic material, represented by a large double-stranded DNA molecule – with the famous double-helix structure. Cells are therefore the fundamental unit of living matter. They take up chemical substances from their environment and transform them. The functions of a cell are subject to regulation, such that the cell acts and interacts in an optimal relationship to its environment. The 'central dogma' of biology describes how information, stored in the DNA, is transformed into proteins via an intermediate product, called RNA. Transcription is the process by which coding regions of DNA (called 'genes') synthesize RNA molecules. This is followed by a process referred to as 'translation', synthesizing proteins using the genetic information in RNA as a template. Most proteins are enzymes and carry out the reactions responsible for the cell's metabolism – the reactions that allow it to process nutrients, to build new cellular material, to grow, and to divide.

Since the 1960s it is known that in fact most basic cellular processes are dynamic, feedback regulated and that cells display *anticipatory* behaviour. In the 1960's, investigating regulatory proteins and the interactions of allosteric enzymes, Francois Jacob and Jaques Monod introduced the distinction between 'structural genes' (coding for proteins) and 'regulatory genes', which control the rate at which structural genes are transcribed. This control of the rate of protein synthesis gave the first indication of such processes being most appropriately viewed as dynamic systems. Figure 1 illustrates the processes of gene expression and regulation in bacterial cells.

Bacterial cells are capable of producing several thousand different proteins, but not all are produced at the same time or in the same amount. The energy consumption for protein synthesis and the relatively short half-life of the RNA molecules are reasons for the cell to control both, the types and amounts of each protein. One example of a global regulatory network is the heat-shock response. When proteins are exposed to extremes of heat, they are said to undergo 'denaturation'. Denaturation is the destruction of the folding properties of a

protein leading (usually) to loss of biological activity. To counteract possible toxic effects from insoluble aggregates in the cell, the change in temperature and the quantity of denatured proteins are 'sensed' by the cell and specific heat-shock proteins are produced. Figure 2 illustrates heat-shock regulation of the DnaK operon in the bacterium *Bacillus subtilis*. The protein DnaK is such a 'chaperon', one of a group of proteins called 'molecular chaperones', which help other proteins to fold properly. These specialist proteins produce barrel like structures, providing an environment for the denatured proteins to refold. The described mechanism is referred to as 'negative control' through repressor deactivation. A repressor protein is a regulatory protein that binds to specific site on the DNA and thereby blocks transcription. New technologies allow us to quantify the activation of genes. For example, DNA microarrays can quantify amounts of RNA produced by the entire genome and also can provide us with time series data. Like other recent developments, this technology requires the availability of genome sequences and is therefore one of the techniques that is said to revolutionize biology in the "post-genome era".

Bacterial microorganisms affect our health in several ways, causing infectious disease (e.g. tuberculosis) as well as providing the basis for medicines such as for example antibiotics produced from *Streptomyces*. Other important biotechnological applications can be found in the agro-food industry. Biotechnology is highly dependent on genetic engineering, the discipline that concerns the artificial manipulation of genes and their products.

Intra- and Inter-Cellular Dynamics: Cellular Weather Forecasting

The previous example, described how do *genes* act and interact within the context of the *cell*. The cells are not running a programme but rather continually sense their environment and make decisions on the basis of that information. To answer how do *cells* act and interact within the context of the *organism* to generate coherent and functional wholes, we need to understand how information is transferred between cells and within cells. Cell signalling or 'signal transduction' is the study of the mechanisms by which this transfer of biological information comes about. Signalling impinges on all aspects of biology, from development to disease. Many diseases, such as for example cancer, involve malfunction of signalling pathways. Downward [2001] provided an excellent account of this field.

Figure 3 illustrates a very basic signalling model. As already indicated in the previous section, bacteria regulate cell metabolism in response to a wide variety of environmental fluctuations, including the heat-shock example above. Therefore, there must be mechanisms by which the cells receive signals from the environment and transmit them to the specific target to be regulated. Receptors are proteins that span the membrane, with a site for binding the signaling compound on the outer surface. Binding of the extracellular signalling compound to the outer surface of the receptor results in an activation of an intracellular protein (the 'response regulator'), for example by phosphorylation. Signalling pathways commonly consist of many more cascaded modules between receptor and genome. There can be numerous intermediate steps before the signal transduction process often ends with a change in the gene expression programme of the cell. In the figure, the phosphorylated response regulator is a DNA binding protein, which serves as a repressor, preventing the RNA polymerase to transcribe the adjacent gene(s).

As well as crosstalk between pathways, negative feedback systems can occur and the time course of a signalling pathway can be critical. It is therefore important to develop experimental techniques that allow quantitative measurements of proteins and protein interactions. Mathematical modelling and simulation in this field has the purpose to help and guide the biologist in designing experiments and generally to establish a conceptual framework in which to think. Illustrations, like those in Figures 1 and 3, are commonly used in biology textbooks research articles. They are extreme simplifications by all standards and knowledge about the

information flow described in them is usually distilled from numerous experiments and generally it is not possible to observe these processes in an engineering fashion (in the sense of “on-line” measurements). However, the illustrations highlight spatial aspects, which in control systems block diagrams usually get lost. The data we currently have available, do not allow parametric systems identification techniques to build predictive models. Instead, it is the systems thinking, the modelling process itself is what often proves useful. A common engineering experience is that we learn most from those models that fail.

Mathematical modeling and simulation is important in many ways, helping to understand the fundamental processes of gene regulation, metabolic pathways, and cell behaviour in cultures. The article by Hasty [2001] provides a survey of such *in numero* molecular biology. The March 2001 special issue of *Chaos* (published by the American Institute of Physics) contains a series of articles on mathematical modeling of gene expression. Smolen [2000] surveys mathematical modeling of transcriptional control and future directions. The signal-oriented approach to cellular models by Kremling et al. [2000] is an example of systems and control theory bridging the gap between cellular biology and metabolic engineering. Progressing from merely descriptive models to predictive models will require an integration of data analysis and mathematical modeling with information stored in biological databases.

Current Research: An International Perspective

In contrast to bio-physical modelling at the molecular level, systems biology develops phenomenological models at the cellular and organismic level and seeks to identify these models from data (e.g. gene interaction networks identified from DNA microarrays). The phenomenological approach defines observables, measurable characteristics, to capture dynamic systems behaviour. It therefore does not explicitly model the physical elements and structure. We shall discuss the limitations of such methodology in detail below, but it is immediately clear that such high-level and behavioural approach has its problems, capturing spatial organisation and its effects, which are vitally important in cellular systems. For computational models at molecular level, Dennis Bray's work [<http://www.zoo.cam.ac.uk/comp-cell/>] provides an excellent example in which pathways are modelled by means of the cell-signalling molecules. The Kaneko Laboratory for Nonlinear Science and Complex Systems [<http://chaos.c.u-tokyo.ac.jp/>] is located somewhere between these two perspectives, combining complex systems theory, developmental biology, biophysics and molecular biology. For example, their mathematical models and simulations of inter-intra-dynamics, describe systems composed of units with internal dynamics and interaction, whose number is assumed to change through the internal dynamics [<http://coe.c.u-tokyo.ac.jp/>].

While the Santa Fe Institute [www.santafe.edu] is most readily associated with the study of complexity in biological systems, the systems approach to biology is in the US pioneered by the Systems Biology Institute of Lee Roy Hood in Washington [www.systemsbiology.org] with more than 150 members of staff. The area of computational systems biology has been introduced, amongst others, by Hirato Kitano in Japan [www.symbio.jst.go.jp], Hamid Bolouri in the UK [strc.herts.ac.uk/bio] and John Doyle at Caltech. In mathematical systems biology, active Centres are, for example, the groups of E.D. Gilles from the Max-Planck-Institute for Dynamic Complex Systems, Magdeburg in Germany [www.mpi-magdeburg.mpg.de] and Olaf Wolkenhauer at UMIST in the UK. The Systems Biology Group at Delft University of Technology [<http://www-ict.its.tudelft.nl/~imds/>] is another example of a group that combines experience in pattern recognition and control theory in the analysis of genetic networks and microarray data. O.Wolkenhauer at UMIST in Manchester coordinates the UK research network in 'Systems Theory and Genomics', funded by the UK Engineering and Physical Sciences Research Council (EPSRC). The German Federal Ministry for Education and Research (BMBF) has launched a £33M Systems Biology Programme in March 2002, and one can expect substantial growth in this area over the coming years. Systems biology provides evidence for a growing involvement

of control engineers in the bio-sciences, not just at the technological level but playing a vital role in the development of novel methodological approaches in mathematical modelling, simulation and data analysis.

The first international conference on systems biology (ICSB) took place in November 2000 in Japan and has since received increasing attention. ICSB'2001 was organised at CALTECH in the US, and is followed by ICSB'2002 in Sweden, in December 2002 [<http://www.ki.se/icsb2002/>].

Large scale simulation projects of cellular processes, cells and the physiome are necessarily relevant to systems biology. These include, amongst others, BioSpice (Berkeley, US) [<http://gobi.lbl.gov/~aparkin/>] the E-Cell Project (Japan) [<http://e-cell.org/>], the V-Cell Project (National Institute of Health, US) [<http://www.nrcam.uchc.edu/>], and the Physiome Project [<http://www.bioeng.auckland.ac.nz/>] in New Zealand.

A number of software tools have been developed to describe various aspects of gene expression and regulation, genetic- and biochemical networks. Depending on which organisational or descriptive level of the biological problem is addressed, these tools are usually not alternatives but complement each other. It is generally recognised that there is no all-in-one package providing a solution but instead a common interface or environment is necessary. The 'Systems Biology Workbench' [<http://bioinformatics.org/sbw/>] and 'Systems Biology Markup Language' [Hucka 2001] are the result of such efforts. The goal is to provide a software infrastructure that helps sharing of simulation software and analysis models. The Systems Biology Markup Language (SBML) is to provide a common XML-based language, which is accepted by a large number of tools. The Systems Biology Workbench on the other hand, uses SBML to transfer models between tools. At present, the focus is on bio-chemical modelling.

Systems Biology: Form and Function

The principle challenge for systems biology is to help the biologist answering the following questions [Sole 2000]:

1. How do *cells* act and interact within the context of the *organism* to generate coherent and functional wholes?
2. How do *genes* act and interact within the context of the *cell* as to bring about structure and function?

More specifically the challenges for mathematicians, physicists and engineers are (see also Figure 4):

1. Dynamic regulation and spatial organisation: the need to capture both, spatial as well as temporal aspects simultaneously (spatio-temporal modelling).
2. Intra- and inter-cellular actions and interactions: the need for large-scale and hybrid-systems modelling and simulation.
3. Crossing organisational levels: from cells, to colonies, tissues, organs and organisms, ...
4. Integrating descriptive levels: genome, transcriptome, proteome, metabolome and the physiome.

5. Combining data analysis and data management: The need to combine computational tools, developed for specific tasks and different organisational and descriptive levels.
6. Relating formal representations (mathematical models, e.g. Boolean networks and rate-equations). Providing a conceptual framework and theoretical foundations for the previous five points.

Gene expression takes place within the context of a cell, between cells, organs and organisms. The inevitable, reductionist approach is to 'isolate' a system, conceptually 'close' it from its environment through the definition of inputs and outputs, we inevitably lose information in this approach. (Conceptual closure amounts to the assumption of constancy for the external factors and the fact that external forces are described as a function of something inside the system). Different levels may require different modelling strategies and ultimately we require a common conceptual framework that integrates different models. For example, differential (mass-action or rate-) equations may provide the most realistic modelling paradigm for a single-gene or single-cell representation but cell-to-cell, and large-scale gene interaction networks could, for example, be represented by logical or finite-state models, using agent-based simulation.

In dynamic systems theory, one would usually ignore spatial aspects. This approach is limited because both, space and time are essential to explain the physical reality of gene expression. The fact that the concepts of space and time have no material embodiment; they are not to be found in the molecules or their DNA sequence; has been an argument against material reductionism. Although this criticism is in principle correct, alternative methods are in short supply. The problem is that although components of cells have a specific location, these locations lack exact coordinates. Without spatial entailment there can be no living cell and for systems biology it is therefore necessary to integrate a topological representation of this organisation.

In his recent book, the biologist Frank Harold gave an excellent discussion of the complexity of cellular processes and provides a compelling argument for the need for more research in complexity studies: "From the chemistry of macromolecules and the reactions that they catalyse, little can be inferred regarding their articulation into physiological functions at the cellular level, and nothing whatever can be said regarding the form of development of these cells. It therefore seems to me self-evident that the quest for the nature of life cannot be conducted exclusively on the biochemist's horizon. We must also inquire how molecules are organised into larger structures, how direction and function and form arise, and how parts are integrated into wholes." [Harold 2001]. This very much reflects Robert Rosen's complaint that "At the moment, biology remains a stubbornly empirical, experimental, observational science. The papers and books that define contemporary biology emanate mainly from laboratories of increasingly exquisite sophistication, authored by virtuosi in the manipulation of laboratory equipment, geared primarily to isolate, manipulate, and characterise minute quantities of matter. Thus contemporary biology simply is what these people do; it is precisely what they say it is." [Rosen 1991]. Rosen mounted the so far most comprehensive attack on Newtonian reductionism and the mechanistic approach to biology. He warned of the naive approach in which biological systems are looked at through the eyes of a physicist. His 'relational biology' is based on the modelling relation and the fact that we try to encode causal entailment in natural systems through formal systems. His ideas are again very relevant to systems biology.

In general, causation is a principle of explanation of change in the realm of matter. In systems biology causation is defined as a (mathematical) relationship, not between material objects, but between changes of states within and between components. Instead of trying to identify genes as causal agents for some function, role, or change in phenotype we relate these observations to sequences of events. In other words, instead of looking for a gene that is the reason, explanation or cause of some phenomenon we seek an explanation in the dynamics

(sequences of events ordered by time) that led to it. "It is systems dynamics, not a genetic program, that gives rise to biological forms and functions." [Harold 2001, p.199]

This 'relational approach' is seeking to identify and explain relationships: interactions of genes, proteins, metabolites and interrelationships, which lead to organisation and structure. Such relational biology is anything but new; it resonates with the philosophy of Immanuel Kant and his 'successor' Arthur Schopenhauer. Linus Pauling observed that "Life is a relationship among molecules and not a property of any molecule." and Henri Poincaré concluded that "The aim of science is not things in themselves but the relations between things; outside these relations there is no reality knowable."

In order to verify theoretical concepts and mathematical models we ought to identify the model from experimental data or at least validate mathematical models with data. In the context of post-genome technologies, the problem of complexity appears then in two disguises:

1. Dimensionality: hundreds or thousands of variables/genes/cells.
2. Uncertainty: small samples (few time points, few replicates), imprecision, noise.

Analysing experimental data we usually rely on assumptions made about the ensemble of samples. A statistical or 'average perspective' may however hide short-term effects that are the cause for a whole sequence of events in a genetic pathway. What in statistical terms is considered an outlier, may just be the phenomenon the biologists is looking for. It is therefore important to compare different methodologies, their implicit assumptions and the consequences for the biological questions asked. To allow reasoning in the presence of uncertainty, we have to be precise about uncertainty. While generally one associates genomics with the availability of vast amounts of data, studying temporal processes, dynamic phenomena, the opposite is true. For example, microarray technology remains either too expensive, or experiments become too complex to generate data that would satisfy a statistical approach to microarray time-course experiments. It is however of paramount importance that we strive to bridge the gap between data and models. In the words of Bertalanffy: "Thus even supposedly unadulterated facts of observation already are interfused with all sorts of conceptual pictures, model concepts, theories or whatever expression you choose. The choice is not whether to remain in the field of data or to theorize; the choice is only between models that are more or less abstract, generalized, near or more remote from direct observation, more or less suitable to represent observed phenomena." [Bertalanffy 1969]

Problems in systems biology are usually approached by a combination of methodologies. The data analyst is required to validate his results with information from biological databases. Microarray gene expression profiles, grouped in clusters using pattern recognition techniques, are often validated by analysing the promoter sequences of the genes involved. The nature of the data and information, obtained at different descriptive and organisational levels are quite different and biological databases in any of these areas are usually not compatible. As the interpretation of data is inextricably linked to databases, an information fusion problem has to be solved. Similar, pattern recognition (for example, used to identify interrelationships among genes) and systems identification techniques used to study gene interactions are usually used in combination and the system scientist should be familiar with both concepts [Wolkenhauer 2001b].

Conclusions

Systems biology and the renewed interest in complexity studies marks a shift away from an often obsessively molecular approach, providing a causal and dynamic account of cellular form and function through a synergy of the bottom-up approach of molecular characterisation with the

top-down perspective of physiology. The characterisation of molecular structures has achieved an impressive accuracy, which at higher levels will be virtually impossible to achieve. Mathematical modelling has long played an important role in biochemistry, biochemical kinetics and metabolic engineering. While structural biology concerns itself with the characterisation of individual molecules, the kinetic or thermodynamic paradigm is that a single molecule behaves stochastically and that deterministic behaviour on the macroscopic scale arises as a statistical property of very many molecules. Modelling relatively simple reactions can lead to sets of unsolvable equations and we often rely on assumptions to describe, for example, how enzymes catalyse (enhance) the transformation of substrates into 'products'. The field of metabolic engineering has taken these models to an industrial scale, researching the targeted improvement of cellular properties or metabolite production via manipulation of specific metabolic or signal transduction pathways. As illustrated in Figure 5, systems biology provides an interface between cell biology, physiology and metabolic engineering. The focus of systems biology is the understanding and explanation of intra- and intercellular dynamics by means of a systems- and signal-oriented approach. It takes account of recent findings in functional genomics, and experimental data generated by post-genome technologies. It is therefore closely allied with the area of bioinformatics, providing information stored in biological databases.

The biggest if not the principle hurdle for complexity studies and systems biology in this exciting period of science is well summarised by Lotfi Zadeh's uncertainty principle:

"As the complexity of a system increases, our ability to make precise and yet significant statements about its behaviour diminishes until a threshold is reached beyond which precision and significance (or relevance) become almost exclusive characteristics."

The need to combine the efforts of biologists, mathematicians, physicists and engineers is reflected in Erwin Schrödinger's observation, true today as it was in 1944: "We feel clearly that we are only now beginning to acquire reliable material for welding together the sum total of all that is known into a whole; but, on the other hand, it has become next to impossible for a single mind fully to command more than a small specialized portion of it." [Schrödinger 1944]

We therefore require scientists that are prepared to invest time and effort into more than one discipline and scientific culture. This will require a change in the education, training and career prospects of interdisciplinary scientists. As John L. Casti observed "To function effectively, the system scientist must know a considerable amount about the natural world AND about mathematics, without being an expert in either field. This is clearly a prescription for career disaster in today's world of ultra-high specialization." [Casti 1992]. As everyone recognises the need to interdisciplinary research, it is important that such a career is not only interesting from a scientific perspective but also a rewarding career option.

One idea to foster and strengthen an multi-disciplinary environment at the life science interface is to establish four year doctoral training programmes. Training centres that provide taught components as part of the PhD training and give students the opportunity to experience various aspects of the bio-sciences in a research environment are a desirable approach.

References

Bak, P. (1997): *How Nature Works: The science of self-organised criticality*. Oxford University Press.

Bertalanffy, L. (1969): *General Systems Theory*. George Baziller Inc.

Casti, J.L. (1992): *Reality Rules: Picturing the world in mathematics*. John Wiley & Sons, Inc.

- Downward, J. (2001): The ins and outs of signalling. *Nature*, Vol. 411, 14 June 2001, 759-762
- Goodwin, B. (1994): *How the Leopard Changed Its Spots: The evolution of complexity*. Princeton University Press.
- Harold, F.M. (2001): *The Way of the Cell: Molecules, Organisms and the Order of Live*. Oxford University Press.
- L.G. Harrison (1993): *Kinetic Theory of Living Pattern*, Cambridge University Press.
- J. Hasty, D. McMillen, F. Isaacs, J.J. Collins (2001): Computational Studies of Gene Regulatory Networks: In Numero Molecular Biology. *Nature Reviews Genetics*, April 2001, Vol. 2, No 4, 268-279.
- Hucka, M. and Finney, A. and H. Sauro, H. Bolouri, J. Doyle, and H. Kitano (2001): The ERATO Systems Biology Workbench: An Integrated Environment for Multiscale and Multitheoretic Simulations in Systems Biology. Chapter 6 in *Foundations of Systems Biology*, H. Kitano (ed.), MIT Press.
- Kauffman, S. (1995): *At Home in the Universe: The Search for Laws of Complexity*. Penguin Books.
- Kitano, H. ed. (2001): *Foundations of Systems Biology*. MIT Press.
- Kitano, H. (2002): An Introduction to Systems Biology. *Science*. Vol. 295, 5 March 2002.
- Kremling, A. and Jahreis, K. and Lengeler, J.W., and Gilles, E.D. (2000): The Organisation of Metabolic Reaction Networks: A Signal-Oriented Approach to Cellular Models. *Metabolic Engineering*, July 2000, Vol. 2, No. 3, 190-200 (11).
- Nature (2000): *Nature Insight – Functional Genomics*, 15 June 2000, Vol. 405, 819 - 846
- Meinhardt, H. (1998): *The Algorithmic Beauty of Sea Shells*. Springer Verlag.
- Rosen, R. (1991): *Life Itself: A Comprehensive Inquiry into the Nature, Origin, and Fabrication of Life*. Columbia University Press.
- Schrödinger, E. (1944): *What is Life?* Canto edition, Cambridge University Press 1992.
- Smolen, P. and Baxter, D.A. and Byrne, J.H. (2000): Modelling Transcriptional Control in Gene Networks – Methods, Recent Results, and Future Directions. *Bulletin of Mathematical Biology*, Vol. 62, 247-292.
- Sole, R. and Goodwin, B. (2000): *Signs of Life: How complexity pervades biology*. Basic Books.
- Weaver, W. (1948): Science and complexity. *American Scientist* 36:536-544.
- Wolkenhauer, O. (2001): Systems Biology: The reincarnation of systems theory applied in biology? *Briefings in Bioinformatics*, Henry Stewart Publications, Vol. 2, No. 3, 258-270.
- Wolkenhauer, O. (2001b): *Data Engineering*. John Wiley & Sons, New York.

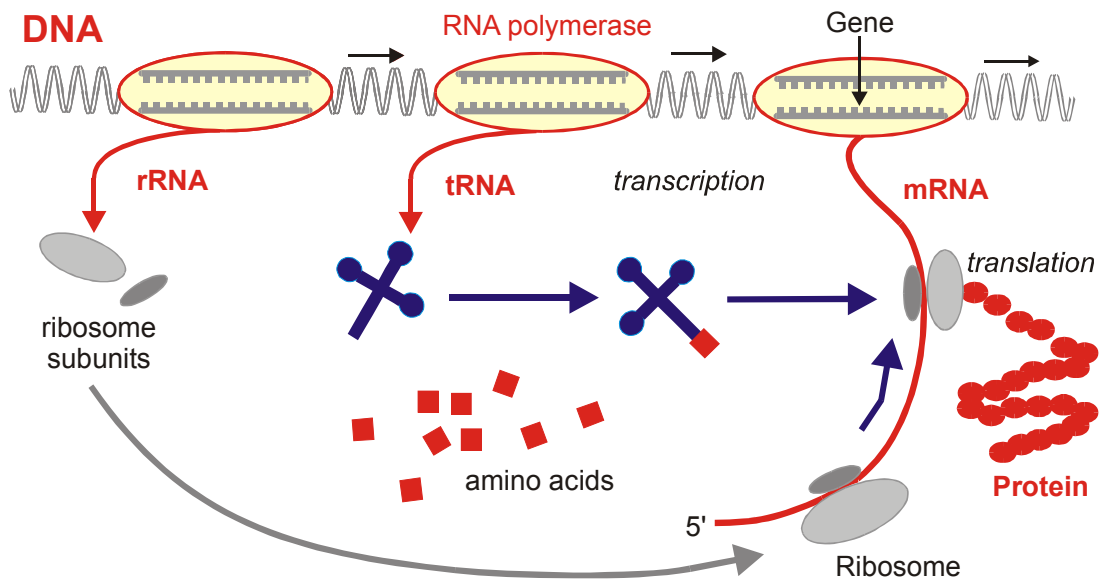


Figure 1: Gene expression and regulation in bacteria. Information, stored in the DNA, is transformed into proteins, via an intermediate product, called mRNA. The short half-life of mRNA and the energy consumption of protein synthesis form the basis for a sophisticated hierarchy of control mechanism.

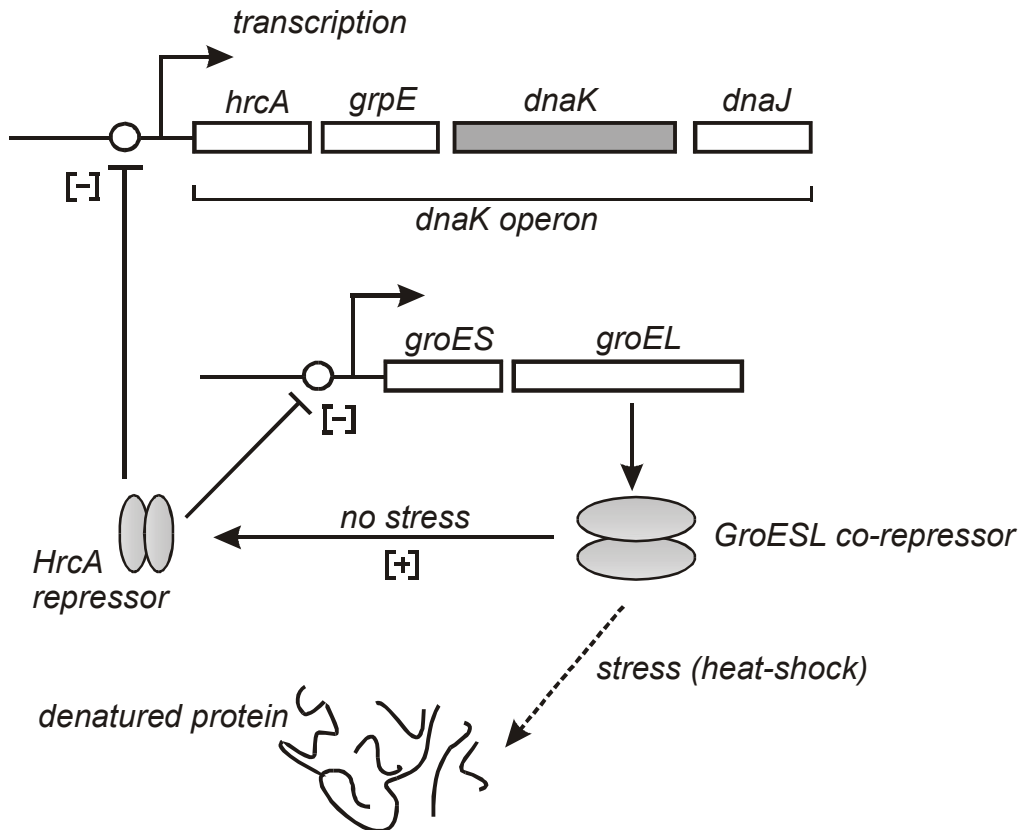


Figure 2: Negative regulation of the *dnaK* and *GroESL* operons in *Bacillus subtilis*. The *HrcA* repressor is a regulatory protein that binds at specific sites on DNA and blocks transcription. In the absence of stress, the *GroESL* co-repressor binds with *HrcA* and thereby increases repression of the *dnaK* operon genes. Upon heat-shock the transcription rate of a group of heat shock proteins, called chaperones, is increased. They build barrel-type structures, which help denatured *HrcA* to refold and regain its function.

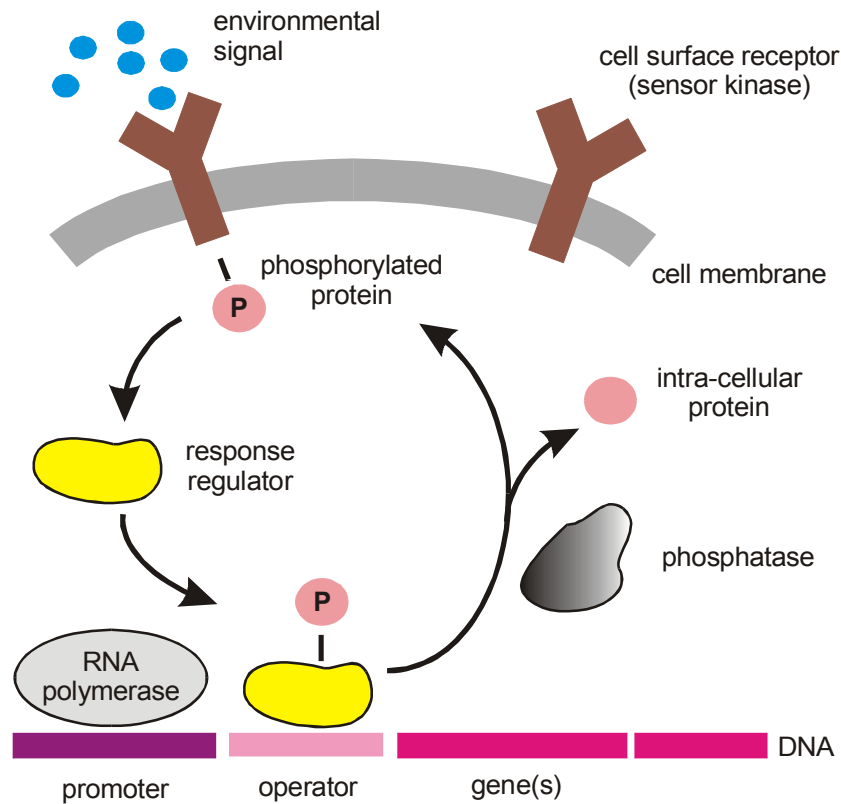


Figure 3: Cell signalling (signal transduction). Intracellular dynamics (gene expression) can be affected by extracellular signals. Receptors, spanning the cell membrane receive signals and transmit the information to activate intracellular proteins (the response regulator). In the picture, the response regulator binds to the operator region of a gene and prevents the RNA polymerase from transcription of the adjacent gene. A phosphatase ensures that the process is continuous.

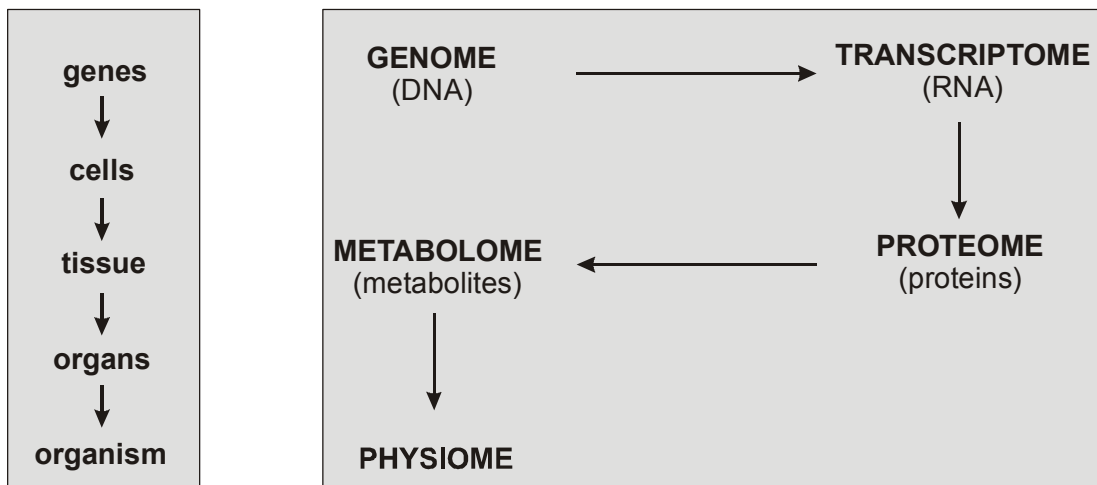


Figure 4: Descriptive levels (right) at which scientists study different aspects of the organisational levels (left – from genes to the physiome). Each descriptive level is associated with particular types of experiments and technologies. The data that are obtained from different levels are different in nature and the information stored in databases is usually not compatible. Nevertheless, for a comprehensive understanding of gene expression and regulation, information from all levels has to be integrated.

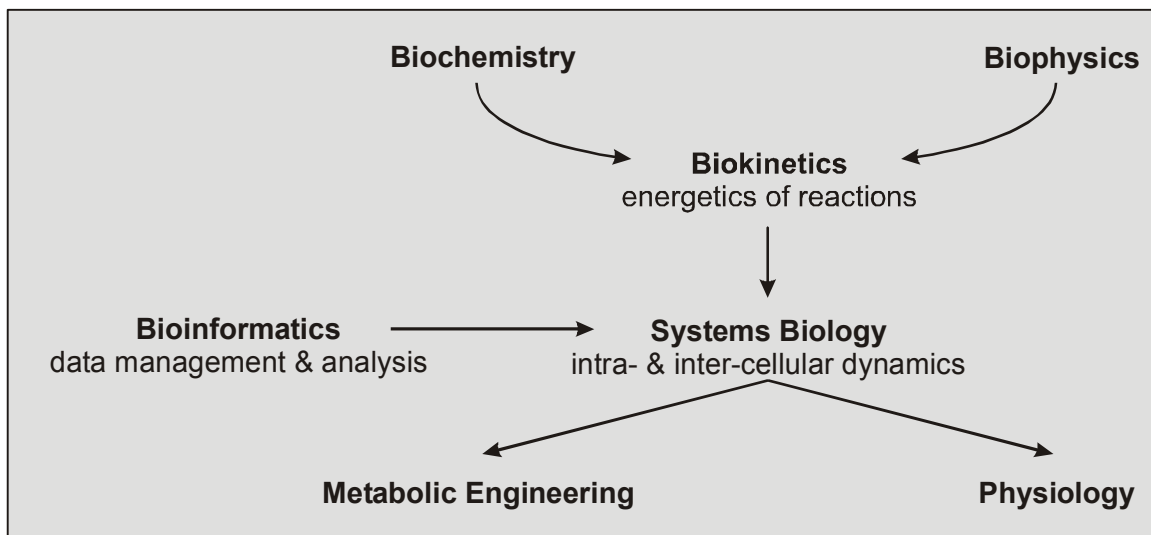


Figure 5: Areas of the biosciences influenced by mathematical modelling. Systems biology provides a link between cell biology, physiology and biotechnological applications. The focus of systems biology is the understanding and explanation of intra- and intercellular dynamics by means of a systems- and signal-oriented approach. It takes account of experimental data and information from biological databases and is therefore closely allied with the area of bioinformatics.