Fuzzy Relational Biology

A Factor-Space Approach to Genome Analysis

Notes, December 2000

Olaf Wolkenhauer

Control Systems Centre

Dept. Biomolecular Sciences and Dept. Electrical Engineering

P.O. Box 88, Manchester M60 1QD, UK

E-mail: o.wolkenhauer@umist.ac.uk

http://www.umist.ac.uk/csc/people/wolkenhauer.htm

Preface

What is referred to as a genomic revolution can be summarised by the following tasks: Analysis of the DNA sequence of a genome, identification of genes, study of gene products (usually proteins). While the identification of components, cataloguing of products and their classification is very exciting, the interrelationships and interactions of these components is what makes organisms like ourselves 'tick'. Once all parts of a clockwork are known, the challenge is to find out how the clock 'works'. New technology makes it possible to study gene activity and allows us to investigate the organisation and control of genetic regulatory pathways. These pathways describe dynamic processes. Their complexity as well as the difficulty in observing them will challenge all analytical tools developed to date. This shift of focus from molecular characterisation to an understanding of functional activity, implies a change from generating hypotheses to testing hypotheses. Problems in genomics will become conceptual as well as empirical.

I will argue that the biggest challenge of bioinformatics is not the volume of data, as commonly stated, but the formal representation of knowledge. The area of bioinformatics has provided an important service to biologists; helping them to visualize molecular structures, analyze sequences, store and manipulate data and information. These activities will continue to be an essential part of bioinformatics, developing working methodologies and tools for biologists. However in order to directly contribute towards a deeper understanding of the biology, bioinformatics has to establish a conceptual framework for the

formal representation of interrelationships and interactions between genes or proteins.

To this date biological knowledge is encoded in scientific texts and diagrams with a noticeable lack of formal mathematical models. There are obvious reasons for this as most physical or engineering systems, for which mathematical modelling has been successful, appear trivial in comparison to molecular or genetic systems and it is not altogether clear whether a mathematical analysis of genomics systems is the long awaited solution. We may however be able to learn in some respect from the engineering sciences. Like biologists, engineers are not born with a love for mathematics and yet they have learned to embrace it as a problem solving strategy or way of thinking. Engineers have been in very much the same situation in which biologists find themselves now: The systems or processes they study, the data they generate are too complex to be dealt with using common sense or intuition. To solve very practical problems, the engineering sciences have therefore learned to translate a given practical problem into a set of (state-space or random-) variables and then to a conceptual framework (such as probability or control theory) to establish relationships among these variables, in order to make predictions, classifications or to influence the system under consideration. Although engineering systems are in most cases rather trivial in comparison to genetic systems, the practise of abstraction for problem solving may in fact prove useful to biologists. It is in this sense that I try to promote systems theory as 'a way of thinking', problem solving strategy supporting the biologists work which remains in many ways of empirical nature.

My interest in genomics is an extension of my childhood career as a Lego-engineer - I enjoyed taking things apart, studying its parts in order to understand *how* things work. This "ontological" quest, however, doesn't really answer the question of *why* things work as they do. For systems that are slightly more complex than a Lego construction, we may find that this question cannot be answered satisfactorily. However, by adding an epistemological question, studying the way we approach the 'how' question, we may be able to go some way towards answering it. I subsequently decided to approach the problem from the bottom up, including philosophical questions into the development of the mathematical model. The following pages are a transcript of an ongoing discussion. The aim is to outline a formal mathematical framework to study gene expression, regulation and function. Apart from a conceptual framework, which we refer to as *fuzzy relational biology*, the objective is to develop a working methodology for the analysis of genomes in terms of gene expression data. With the presented approach, we are looking for ways which characterise molecular systems in a general way, and quite independently of their physical or chemical constitution; we are seeking a theory of 'principles'. The representation of biological knowledge (concepts, facts, and rules) is based on fuzzy mathematics and systems theory.

For many years I have had the wish to eventually merge my interests in mathematical modelling, data analysis, the philosophy of science and biology. The progress of molecular biology in recent years has meant that there is an increasing need for interdisciplinary research and it has become possible to make a passion part of my work at University. In his famous essay "What is life?" in 1944, Erwin Schrödinger's provides motivation for interdisciplinary research which is again valid today [66] :

> *"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 specialised portion of it."*

Like Schrödinger - a physicist writing about biology, the biologist Jaques Monod, taking an interest in philosophy, was worried about risk one takes when crossing boundaries and going public. I do not hesitate to admit my limited knowledge in the biology and mathematics I am going to discuss. Great minds like Schopenhauer, Popper, Einstein, Bohm and Rosen however demonstrate that if we are not studying the wider context of our work or consider its consequences, we miss out on the excitement and fulfilment such work can give. Their books have been of constant inspiration and encouragement.

OLAF WOLKENHAUER

Manchester June 5, 2003

Acknowledgments

I have benefited from an invitation by Robert Babuska to work at the Technical University Delft in the Netherlands from November 1999 to June 2000. Large parts of the document were prepared in Delft, where the control laboratory and bioinformatics group provided an ideal environment and atmosphere for such research. In times of tight departmental budgets, UMIST tolerated my obsession with books. The work arose from reading a number of books over several years. My own views slowly emerged from these studies as a synthesis of the work of my favourite authors. Considering the scope and nature of my passion, it seemed impossible to have someone I could discuss my views with. Allan Muir is one of the rare species that is prepared to listen to a mixture of math, philosophy, politics and biology and can at the same time provide an inspiring and thought provoking response. His comments during a number of discussions are much appreciated. He is lucky to live away from Manchester as otherwise I could not have resisted to bother him more frequently.

Over the last four years, my research has been supported by the UK Engineering and Physical Sciences Research Council (EPSRC), Biotechnology and Biological Sciences Research Council (BBSRC). In preparing this book I have drawn extensively upon existing literature. In particular the work of Robert Rosen and the book by H.X. Li and V.C. Yen has had substantial influence. Finally, for a number of years I am using MiKTEX, a free LATEXinstallation provided by Christian Schenk, who indirectly has had a considerable supportive influence on my work.

O.W.

Contents

Introduction

The work, out of which the present document evolved, began with a review of current and past approaches to quantitative and formal modelling of processes in organisms. This lead to a number of deliberately bold claims and critical remarks, in the hope to emerge with a novel approach to the most exciting problems available: "how do organisms function?" and "how do we know?".

The world as we experience it is representation and as such accessible to science through ordinary perceptual or sensual experience and is usually described in terms of individual material objects (e.g. the DNA molecule) and abstract objects or concepts (e.g. genes, gene function) which can be investigated scientifically. Our 'experience' is realised through observation and measurement in a scientific experiment and to make *a priori* discoveries, i.e., predictions about the nature of this world of objects, we must renounce the attempt to know what they are in themselves. Objects are representations for the subject and we can only have knowledge of empirical objects using the a priori forms of space, time, and causality. In the present text we carry this philosophical position over into a conceptual framework and working methodology for genomics. Studying genetic systems, we therefore try to avoid ontological questions and instead provide a phenomenological model of gene expression, gene function and interactions.

In physics, the assumption that an apparently complex natural systems can be explained by a simple formal model which is general or universal has

been successful. For the physicist, at a close look, biological systems are structurally and functionally determined by basic physical laws. Biologists describe instances of complex systems that suggest simple mechanisms. Should we therefore consider biological concepts as instances of the theories, principles, and concepts of physics?

Measurement or observation provide the basis for any scientific approach. An observed regularity (order, invariance) is seen as confirmation for the existence of some kind of law or principle by which the natural system under consideration can be explained. Philosophy has shown that the formulation of causal entailment in space and time is the principle mechanism of human perception and conception. Can we ever find an accurate, true, unique and unambiguous explanation?

The Newtonian paradigm, in which phenomena are described by states and transitions between them, fails to describe aspects of organisms that do not possess a single, universal principle. Consequently, an organism can present itself to different observers in various ways. Therefore, rather than to speculate about the principles of inner structures, organisation and the resulting behaviour, it seems useful to devise an approach that starts with the observation of molecular and genetic systems to develop phenomenological models. Requirements or expectations for such a framework are:

- ▷ to fit experimental data, allowing predictions,
- ▷ to have explanatory value in testing hypotheses,
- ▷ to allow conclusions about which variables should be measured and why.

The text is organised as follows. In the first section, the study of biological phenomena is introduced as a constructivist activity ("Zooming In"). A minimalist introduction to molecular biology outlines the context of this text ("The Big Picture"). The following main part proposes a novel formal approach to genomic analysis ("Taking a Snapshot"). In "Image Analysis" we introduce the concept of approximate reasoning to analysis composite phenomena. In order to validate and investigate the concept in greater detail, a 'project plan' is provided ("Image Enhancements"). Finally, related and previous work is discussed ("Moving Pictures"). A summary of mathematical symbols and notation is provided in the appendix.

The text as a whole should be understood as a proposal rather than a fixed result. As yet there are a number of holes which to fill will hopefully I can enjoy for some years to come. Modesty is also required as for any of the holes, we may find the whole venture in danger.

I.1 A CASE FOR MATHEMATICAL MODELLING

There is no science without theory. Any investigation necessarily takes place within a contextual framework. The design of experiments, relies on assumptions and provides choices, as does the (statistical) analysis of data. A mathematical model, whether derived from experience or identified from experimental data, requires us to work within constraints and limits. These pre-conceptions, assumptions, choices are in fact the essence of a 'theory', the experiment, the statistical and mathematical model are only means to further develop or validate the theory. As Henri Poincare pointed out 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." Science, aided by mathematics, is the most rational and objective approach to explain the natural world and yet, throughout this book, we shall find that mathematical modelling is an example of art within science.

Studying natural systems, any scientific approach relies on modelling (analytical, numerical, or observational). In case a mathematical model is an *appropriate* representation of the natural system under consideration, it allows us to make *predictions* about specific values or provides *explanations* by showing that certain things follow necessarily from others. A model is accepted or validated by evaluating its accuracy, i.e., how well the formal system describes the natural system. This can be done by matching experimental observations and/or measurements with the theory. No doubt, if the predictions are accurate, we can apply the model in many useful ways but what does the structure of the model tell us about the principles by which the natural system functions? Usually, there exist a number of alternative models which quantitatively are equally valid but yet are very different in their internal construction or are arrived at in a number of ways (depending on which 'school of thought' is adopted). As a result, we may find that the explanation to what principles are responsible for the observable pattern or data can differ. Therefore even with formal mathematical models, suggesting on the outset rigor and certainty, a semantical problem remains and any ontological question of "How does nature works?", is accompanied with an epistemological problem of "How do we know?". Only if we are approaching both questions together we are able to answer questions regarding the nature of reality.

Successful mathematical modelling requires an awareness of alternative approaches, with equal importance put to synthesis: finding similarities between models that appear different, and analysis: identifying differences between models which appear similar. I would summarise the process of mathematical modelling as follows. Beginning with observations, we start with a question or hypothesis which is investigated within a conceptual framework. This will eventually form the basis for a theory. The latter is tested by validating its model(s) with experimental data (and/or observations). The power of math-

ematical modelling lies in the fact that it is a form of abstraction; it leads to generality. In other words, a good model does not only explain one specific set of data but a class of similar data or systems. Analytical models, leading to quantitative predictions and numerical models, using simulations, provide us with data and/or observations to test whether our methodology is working. Using parametric models, our objective is to identify the system parameters directly from data.

In order to make a mathematical model feasible, assumptions are necessary. We may consider the process as dynamic but time-invariant or even as static if a dynamic formulation is too complex. Relationships between variables of interest are often assumed to be linear and other variables may be taken as independent or unrelated even when they are in fact linked. We frequently ignore variables to obtain simpler models. To be able to identify system parameters from data, further assumption on the quantity and quality of the data are necessary. Given these assumptions are *acceptable*, a number of methodologies are available to obtain precise and yet general models. These techniques are the result of decades of research in system theory, i.e., cybernetics, control theory, time-series analysis and so forth. The systems considered in modern life-sciences provide for these efforts their biggest challenge. The complexity of most molecular and genetic systems, such as regulatory processes of gene and protein interactions, forbid any conventional approach without considerations of uncertainty in data, in measurement, in modelling and in the representation of an analytical model in a computer. To allow reasoning in the presence of uncertainty, our modelling approach has to be precise about uncertainty in order to be of value to the scientist. Modelling uncertainty, i.e., imprecision, randomness, ambiguity, vagueness, and fuzziness is of vital importance for mathematical modelling to succeed in this area.

Functional genomics has leapt from a futuristic concept in the 1980s to an established field of biological research. Genome sequencing projects have lead to an inventory of genes for which we can now measure activity levels (mRNA abundance - gene expression) and protein interactions. With the availability of data on the molecular level, we can not only assign a function (role) to the identified genes but also investigate the organisation and control of genetic pathways which make up the physiology of an organism. It is increasingly appreciated that we can't understand cells by taking them apart piece by piece, since their biochemical pathways form tangled networks. Rather than studying individual components and products as what they are in themselves, we ought to find models of the interactions; the focus is to change from molecular characterisation to an understanding of functional activity through mathematical representations of gene or/and protein interactions. Proteomics research shows that most proteins interact with several other proteins. The classical view of protein function, focused on a single protein molecule, de-

scribed the action of a protein as a catalyst of reactions or its binding to another molecule. With new experimental techniques and data available, it appears appropriate to describe the function of a protein in the context of its interactions with other proteins in the cell. Some of the functional linkages reflect metabolic or signalling pathways; other linkages reflect the formation of complexes of macromolecules.

With the emphasis on functional linkages, and relationships between network variables, problems in proteomics and genomics are increasingly of conceptual nature rather than of empirical nature[1]. With the need to capture various types of uncertainty in modelling, the "fuzzy relational biology" we develop is based on sets of objects, (equivalence and similarity) relations on these sets, and the linkage between factors evaluating the objects in different experimental contexts. Therefore, although the mathematics of the conceptual framework is abstract and initially suggests little relation to the empirical problems of the biologists, the assertion is that if we succeed with its foundations, the models will have more practical value than those provided by classical system theory (successfully applied in many engineering problems). The most important requirements for a modelling framework are that it is precise (honest) about uncertainty, that we can quantify its accuracy and hence allow comparisons between models, and that it can tell us something about which variables to measure and why. Thus, to face the challenges in proteomics and genomics, there is a need for more sophisticated knowledge representations (system models) that interpret the data, organise facts, observations, relationships and even hypotheses that form the basis of our current scientific understanding.

We see an ever-increasing move toward inter- and trans- disciplinary attacks upon problems in the life-sciences and a new mathematical biology is emerging. The system scientist has a central role to play in this new order, and that role is to first of all understand ways and means of how to encode the natural world into 'good' formal structures. System theory is not a collection of facts but rather a way of thinking, and the modelling process itself may be more important than the obtained model. It is concerned with the study of organisation and behaviour *per se* and often it is when the models fail that we learn the most. As the mathematician David Hilbert once said, there is nothing more practical than a good theory. Building on experimental data from cells, we use the power of analytical and numerical methods to explore gene expression, gene function and gene interactions. We are well aware of the fact that it is risky to construct mathematical models too remote from

[1]The term *empirical* means based or acting on observation or experiment, not on theory; deriving knowledge from experience. An empirical approach implies heuristics. The term *heuristic* means allowing or assisting to discover, proceeding to a solution by trial and error.

the process they are supposed to emulate. Models should not only be used to explain data but ought to be identified from data. We shall try to bridge this apparently impossible gap between theory and practice, models and data.

Biology is not reducible to physics and I hope biology is not adopting a mechanistic philosophy. The enormous diversity and complexity of things found in the natural world, both in common experience and in scientific research, cannot be reduced to nothing more than the effects of some limited framework of principles such as those governing machines or automatons. The history of science has continually contradicted the philosophy of mechanism[2] and to talk of 'computations in cells' is misguided and an excessively simple representation of reality.

[2]For a more substantial critique of the philosophy of automaton, see David Bohm's *Causality and Chance in Modern Physics*, Routledge 1957 (1997).

1

The Big Picture

In this chapter we briefly introduce the biological background to the remaining part of the text.

1.1 ZOOMING IN

Energy or matter appears to have been the primary object of science. Its study in the *phenomenal world* is based on changes and for anything to be different from anything else, either space or time has to be presupposed, or both. We shall adopt Schopenhauer's philosophy in which changes in space and time are the essence of *causal entailment*. The subjective correlative of matter or causality, for the two are one and the same, is the *understanding*. "To know causality is the sole function of the understanding and its only power. Conversely, all causality, hence all matter, and consequently the whole of reality, is only for the understanding, through the understanding, in the understanding" [38].

Aspects of the phenomenal world can be examined at different scales. Analysis at any given level of magnification may be successful for solving some problems and not for others. In Figure 1.1, levels of magnification and their associated fields of investigation are illustrated. We shall contend that scientific theories deal with *concepts*, not with reality. All theoretical results are derived from certain formal assumptions (axioms) in a deductive manner. In the biological sciences, as in the physical sciences, the theories are formulated

as to correspond in some useful sense to the real world, whatever that may mean.

Classical *genetics* was founded by Gregor Mendel in the 1860s as the science of heredity. Since then, biologists have "zoomed in" to study cellular processes and structural properties, creating the specialist area of *cell biology*. In the 1950s James Watson and Francis Crick described the three-dimensional structure of DNA as thus founded *molecular biology* which since then has explored the biochemistry and physiology of macromolecules such as the DNA and its immediate products. With the knowledge of sequences and structure, accumulated over the last few decades, a new era dawns with *genomics* and *proteomics*. It is again time to "zoom out", to consider technologies, taking 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, ultimatively, the organism. Genomics is the science that studies the link between proteins and genes. Proteins are cellular effectors, serving such roles as enzymes, hormones and structural components. Protein are usually represented as a linear polymer composed of building blocks known as amino acids. The template for a protein is stored in a molecule known as DNA. Modelling the entire amount of genetic information in a human (known as the genome) is a humbling exercise – it contains about 80 000 genes (organised into large structures known as *chromosomes*). In addition to the regions of DNA which code directly for proteins, much of the genome sequence is regulatory, or serves other purposes such as physical storage and others yet to be discovered. It is however anticipated that the use of model organisms will lead to functional assignment of most of the human genes. A structural hierarchy of genome information is shown in Figure 1.2. Beginning with the organism, consisting of cells, the genome is stored as the DNA molecule in chromosomes. Up to this point, all elements are physical or material objects.

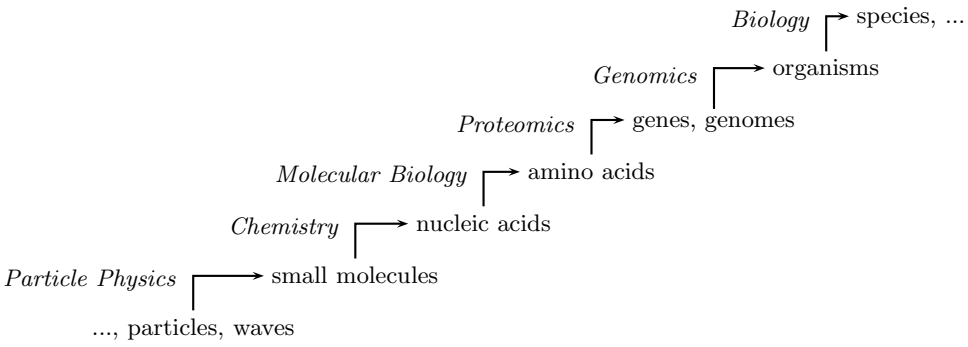


Fig. 1.1 Levels of magnification in the study of biological phenomena.

Hereafter we consider biological phenomena as specified by *genomes* containing *biological information* to construct and maintain an organism and its functions. At present, the dominant approach to describe cellular life forms is based on *DNA* – a molecule made up of a sequence of *nucleotides* whose four possible bases are denoted by the letters A, C, G, T. The information captured by a gene is usually described as being read by *proteins* that attach to the genome at the appropriate positions and initiate a series of biochemical reactions referred to as *gene expression*. Although, *genes* have been seen as subsections of the genome sequence which contain biological information, we shall here describe genes (and indeed genomes) not as structural or physical entities but as *concepts*. Although empirical methods have been the dominating scientific method in life sciences, in the present text we are going to promote a conceptual formal approach. Two particular aspects of a concept are going to play a major role: the *intension* and *extension* of a concept. The intension of a concept is defined by a set of properties and relations subsumed or synthesised by the concept, while the extension of a concept is defined by the set of all objects to which the concept applies.

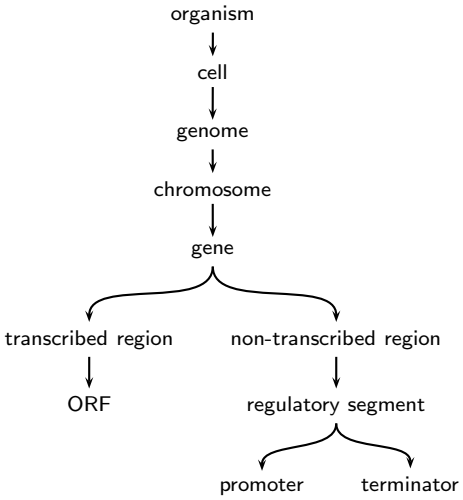


Fig. 1.2 Hierarchy of structural objects in genomic analysis.

Current analysis of genomes is driven by the prediction of functional features at the molecular and cellular level; it is commonly based on the presence and absence of certain genes in the context of phenotypic expectations. Information about gene transfer, the loss, acquisition or displacement of pathways and the correlations of gene occurrences enables biologists to identify functional properties. Our aim is to formulate a mathematical, conceptual framework to represent genes and gene expression. Using this formal language the

objective is to build models for gene-interactions. The approach is purely *relational*, i.e., it makes no reference to the material elements out of which an object or process under consideration is composed. The present text therefore looks at molecular biology as through the glasses of a *system scientist*[1]. Depending on the reader's perspective or views, suspicions of short-sightedness or colour-blindness would be expected and should be communicated to the author.

Although many sciences continue to rely on concepts and methods that commonly derive from the physical sciences, it is now generally appreciated that basic physical concepts such as energy, linear models and reductionism, fail for a wide range of natural systems. They fail to capture the dynamical richness of large, nonlinear and strongly interacting processes and the present work has been motivated by these problems, working towards a more integrative biology. The mathematics of this proposal is rooted in fuzzy mathematics and systems theory which I have described in [77] and [78] respectively[2]. We understand *system theory* as the theory of formal mathematical models[3] of real life (or conceptual) systems. Rosen [61] described system theory[4] as the study of *organisation* per se. We also adopt the definition of a system by Klir [33] who describes a system as a set of *objects* associated with *relations* defined on the set of objects. As such, a system is represented by a subset of some Cartesian product of given sets. We shall distinguish between two systems in particular: a *natural system* – a particular biological process or concept under consideration and a *formal system* or mathematical *model*. Establishing a modelling relation between these two systems is the very definition of a scientific investigation. Properties of formal models can be investigated either by mathematical deduction or by computer simulation. From a biologists point of view, biological phenomena are studied in terms of *informal* aspects, dealing with the meaning, interpretations, significance, objectives, values, and so forth. On the other hand, system theory deals with *formal* aspects of observations, i.e., the form (structure) in which the relationship between the attributes appear.

[1]It should be noted that the various examples taken from biology, used to illustrate a formal concept, are not supposed to compose and demonstrate a single particular case for which the factor space approach is developed.
[2]The help readers, less familiar with mathematical notation, the appendix provides a summary notations and symbols used throughout the text.
[3]The meaning of the term *model* is that of some created thing or process which behaves similar to another. The purpose of a model is to enable inference and hence to extend our knowledge. The underlying assumption for the present text is that mathematical models are valuable to organise data, to study interactions in complex systems, and to understand essential features of biological systems which otherwise would be difficult to achieve.
[4]Probably the most comprehensive treatment of systems theory, its philosophical and formal foundation and application to various areas is the book by Klir [33].

Before we devise a model or conceptual framework, we should clarify their purpose. Therefore, why do we make system models?

- ▷ .. to organise disparate information into a coherent whole.
- ▷ .. for a logical/rational analysis of interactions and dynamics.
- ▷ .. for decision making, i.e., prediction, classification and control.

Formal models are useful in verifying and correcting conventional wisdom and intuition which are hampered by the limitations of our perceptual and conceptual abilities. Cognitive science has provided us with many examples of difficulties that common sense and intuition have with phenomena outside the scope of common or everyday experience. Mathematical models, in particular those which fail, can be useful in complementing the scientist's endeavours to describe/define knowledge in his field. In molecular biology, and the biotechnology it has created, mathematical models have been primarily used in the areas of metabolic or biochemical engineering. Although the objectives of metabolic engineering[5] are related to those in functional genomics [3], biologists to this date rely largely on 'mental models' based usually on empirical, often heuristic methods. The area of bioinformatics has been increasingly important to biologists, helping them to extracting pattern from data, which they subsequently turn into biological facts and knowledge. More recently, discussions on the future of the life sciences suggest a need for mathematical models to go beyond their current status and to provide a rigorous, systematic, and quantitative interface between molecular processes and macroscopic phenomena. In other words, new mathematics and novel methodologies are required to contribute to the conceptual or theoretical framework in which biologists study organisms. We return to a discussion of these matters in Section 3.4, when we suggest system theory as a way of thinking in this direction.

For what follows, the following texts are the main references. Most of the formal methods and fuzzy mathematics is discussed in [78]. With regard to biology, the book by Brown [8] serves as the main reference while for factor spaces most ideas are adopted from Wang [73] and Li [37]. The latter also provides an introduction to 'factor space theory', introduced by Peizhuang Wang in 1981. We use factor space theory because it provides a general but formal framework to represent knowledge. We shall describe biological knowledge in terms of *objects*, *concepts* and *rules*. This is based on the view that

[5]To this date metabolic engineering is largely based on linear systems theory and consequently relies on a number of assumptions that are often difficult to justify for biological systems. The difficulties are in particular that in a Newtonian framework, from which these approaches are derived, the system is often assumed to be closed – can be disconnected from its environment, non-linearities are assumed negligible. Multiple steady states and time variant dynamics increase the complexity of these approaches. Perturbations to cells induce a multi-gene, multi-transcript, multi-protein response which is difficult to capture with conventional reductionistic techniques.

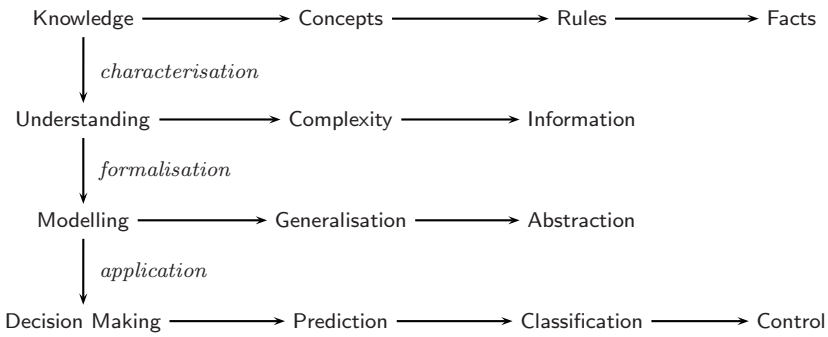


Fig. 1.3 Conceptual outline of mathematical and experimental genome analysis.

organisms are *organised* natural systems and organisation inherently involves *function*. The aim of a system theoretic approach is to provide a *relational* description of a molecular or genomic system which can be matched with observations (data). Hence we favour Rosen's relational approach [61] to describe the function and behaviour of natural systems. Such formalisation, motivated by generalisation and abstraction leads to what is called a formal system or model. In Section 2, we are going to describe a complete philosophical framework which forms the basis for the system theory and fuzzy relational biology developed thereafter.

1.2 MOLECULAR BIOLOGY

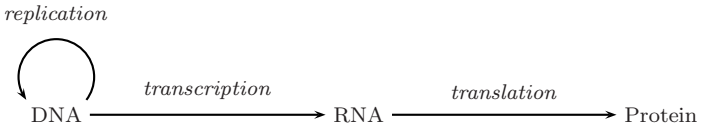
The purpose of this section is to provide a brief, simplified overview of the molecular biology required for subsequent sections. The introduction of the terminology follows the currently common practise in textbooks.

The basic unit of life is the cell, an organised set of chemical reactions bounded by a membrane. Organisms are large collections of cells working together, with each cell having its own identity and function. The large molecules of a cell are of three types: DNA, RNA, and proteins. These macromolecules are made by joining certain small molecules together in polymers[6]. While the DNA polymer is made up of four nucleotides with bases $\{A, C, G, T\}$[7], RNA is a sequence of ribonucleotides, represented by four letters $\{A, C, G, U\}$. Both DNA and RNA are nucleic acids. Proteins are also polymers and here sequences are constructed from an alphabet of 20 amino

[6] A *polymer* is a compound made up of a long chain of identical or similar units.
[7] DNA stands for deoxyribo nucleic acid; RNA for ribonucleic acid. Bases Adenine, Guanine, Thymine, Cytosine and Uracil are abbreviated by their first letter.

acids. A *gene* is a DNA segment containing biological information for the manufacture of proteins – coding for an RNA and/or polypeptide[8] molecule. Sequences for these proteins have directionality as well as a characteristic 3-dimensional structure.

Under gene expression, we understand the series of events by which the biological information, carried by a gene, is released and made available to the cell. This process is often referred to as the *central dogma* of molecular biology, describing the information flow for organisms with DNA genomes as follows :



The arrows in the diagram describe the synthesis of new macromolecules guided by the sequence of an existing macromolecule. The information necessary to control the chemistry of a cell is stored in the DNA macromolecule. The transcription step involves the separation of the double-stranded DNA polynucleotide chains, each serving as a template for the synthesis of a complementary RNA strand. In transcription, the same base-pairing rules apply as in DNA, except that uracil (U), which occurs in RNA instead of thymine (T), pairs with adenine (A). Hence, the formed RNA strand carries the same genetic information as the DNA strand. If the RNA, transcribed by the DNA, codes for protein then it is called *messenger RNA*, mRNA for short. Messenger RNA passes from the cell nucleus to the ribosomes[9] in the cytoplasm[10] which are the sites of protein synthesis. The mRNA attached to ribosomes serves as a template for the information in the polypeptide chain. At this step the rate of protein synthesis is related to the quantity of functional mRNA available which in turn depends on the rate of transcription of DNA to RNA, that is, on the rate of delivery of mRNA from the nucleus to the cytoplasm, and on the rate of mRNA degradation. This short description and the illustration in Figure 1.4 can only hint at the complexity involved in the *regulation* of gene-expression. An important point however for subsequent development of a formal approach to the study of gene expression is that for a comprehensive picture of gene expression and function we require information from various levels and any formal model should have the capability to integrate observations/measurements from the transcriptome and proteome level. Not only that experimental techniques for the proteome and transcriptome are quite different, the data are of different type and are stored in various formats. An *information fusion* problem occurs at this point.

[8] A *polypeptide* is a polymer of amino acids.
[9] *Ribosomes* are ball-like structures that act as 'workbenches' for making proteins.
[10] The *cytoplasm* is the interior portion of the cell exclusive of the nucleus.

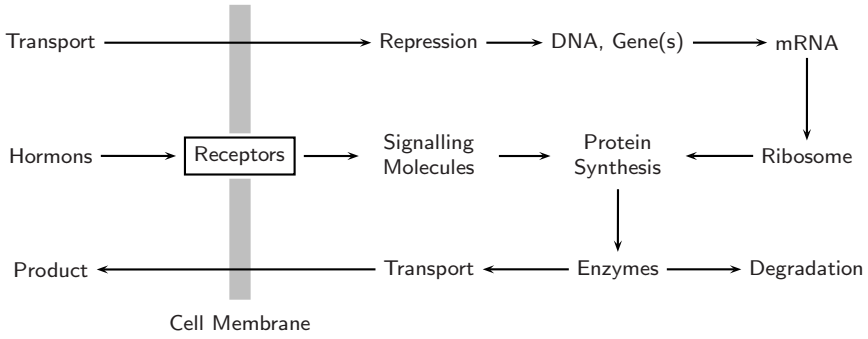


Fig. 1.4 Trivialisation of the regulation of gene expression.

Many genes in eukaryotes[11] (organisms with a membrane bound cell nucleus) are split into *introns* and *exons*. While an exon is a region that encodes a portion of a protein, introns are regions of RNA that do not (appear to) contribute information for the formation of protein. Consequently, in mRNA the introns are spliced out such that exons join up. The rules that determine which triplet of nucleotides codes for which amino acid during protein synthesis is called *genetic code*. The part of a protein-coding gene that is translated into protein is called the *open reading frame*, ORF. Each triplet of nucleotides in the ORF is a *codon* that specifies an amino acid in accordance with the genetic code. The regions immediately after the start of transcription are known to influence gene expression. In other words, gene sequences allow the identification of ORFs which, with increasing certainty, allow for functional assignment. The majority of assigned ORFs relate to metabolic functions.

In the near future, sequencing projects will provide complete genome sequences for only a relatively small number of organisms. This is particularly true of eukaryotes, where only few complete genome sequences exist in the public domain (e.g., yeast: *Sacchoromyces cerevisiae* and the nematode worm, *Caenorhabditis elegans*). Thus for the foreseeable future, it will be necessary to exploit the information available from the relatively few complete genome maps to study gene structure and function in those organisms for which only fragmentary sequence data are available [48, 8]. Such data are produced both from function directed sequencing and from sequence programmes such as shot-gun[12] libraries (which do not intend to obtain closure to assemble a complete genome sequence). Random clone sequencing provides specific gene/ORF sequence data as well as information on synteny which we will here consider as 'data objects'. The post-genome challenge

[11] An *eukaryote* is an organism in which the genetic material is localised in a membrane-bound compartment called the nucleus. Eukaryotes include animals, plants, and fungi. *Prokaryotes* on the other hand lack a proper nucleus. An example are bacteria.
[12] The *shotgun* approach is a genome sequencing technique in which molecules are randomly fragmented and subsequently analysed.

is to be able to interpret and use the genome data: focus is shifting from molecular characterisation to understanding functional activity. The identification of patterns and prediction of properties in metabolic[13], regulatory, or developmental pathways has applications in biotechnology. In addition to gene-sequence data, large-scale RNA assays and gene-expression microarray studies become increasingly important in functional genomics. Microarrays (gene chips) will be further discussed in Section 1.2.2. They can be used to study gene-expression; to study which gene products are made, how much is made and under what circumstances [85].

Once a DNA sequence is available, various methods can be employed to locate genes by

- searching DNA sequences for special features associated with genes,
- experimental analysis of DNA, searching for expressed sequences.

The function of a gene can be assessed by

- homology analysis,
- determining the effect its inactivation has on the phenotype[14] of the organism.

Genes are non-random sequences[15] of nucleotides that code for proteins and can be identified by open reading frames (ORFs) consisting of a series of codons that specify the amino acid sequence of the protein that the gene codes for. The ORFs begin and end with characteristic codons. With three nucleotides forming a codon we have, for double stranded DNA, six possible reading frames and the success of ORF scanning depends on the frequency with which termination triplets appear in the DNA sequence. Simple ORF scanning is less effective for higher eukaryotic DNA with more space between real genes. We can improve procedures for ORF scanning by taking account of *codon bias* – the fact that not all codons are used equally frequently in the genes of a particular organism. Other features to identify genes are for *CpG* islands – upstream regions of vertebrate genes in which the GC content is greater than average. Such *measurable* or *quantifiable* features or *factors*

[13]To this end *metabolism* is viewed as a collection of biochemical reactions, and a *metabolic pathway* is a connected series of these.
[14]The *phenotype* of an organism are its forms and behaviour, i.e., observable characteristics. The *genotype* on the other hand is a description of the genetic composition of an organism, i.e., the causal basic for the phenotype. The study of the genotype-phenotype dualism is the context in which biologists study the function of gene via measurements of gene expression.
[15]Although genes display pattern within a genome they are not simply a sub-sequence of the genome but should be considered as a 'functional unit'. The fact that it is not easy to define a gene as a structural element motivates our formal definition (Section 3.1) of it as a *concept* characterised in terms of *factors*.

form one component of the mathematical model introduced further below. The second important source of information we integrate in our formal model is homology analysis. This information is often derived from experiments and expressed *qualitatively* as an *observation*.

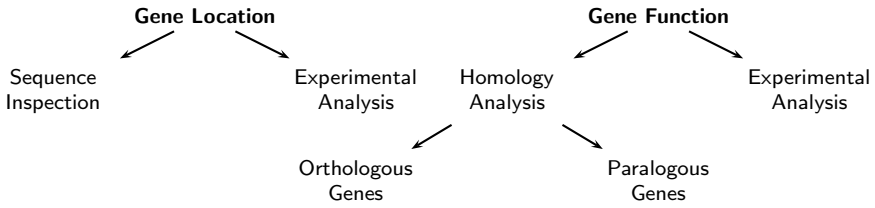


Fig. 1.5 Conventional techniques used to determine gene location and function.

Once a gene has been located, the next question is what function[16] it has. In homology analysis we use the fact that homologous genes are ones that share a common evolutionary ancestor, revealed by sequence similarities between the genes. To identify functions in an unknown gene one focusses on *orthologous* genes which are present in different organisms and whose common ancestor predates the split between the species [26]. A test for homology can be carried out with well established software tools such as BLAST [Altschul et al 1997]. Although the decision is non-fuzzy, that is, genes are either evolutionary related or not, the information obtained from a database search is associated with a likelihood. As before with gene location, in addition to a computerised analysis of sequence data, gene function can be assigned by experimental analysis. One such method is to study the phenotypic effect of gene inactivation.

An important part to the understanding of genomes will be a family of technologies called *genomics* that study the process 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 the cell, tissue and, ultimatively, the organism. It is evident that similarities between homologous

[16]The term *gene function* is ambiguous. Gene function associated with open reading frames frequently refers to functions on a biochemical level. Gene function as an influence of the gene product on the phenotype is not directly related to ORFs. Gene function can be determined by studying gene expression. Gene Expression is the process by which a gene's coded information is converted into the structures present and operating in the cell. Expressed genes include those that are transcribed into mRNA and then translated into protein and those that are transcribed into RNA but not translated into protein (e.g., transfer and ribosomal RNAs). A gene product is the biochemical material, either RNA or protein, resulting from expression of a gene. The amount of gene product is used to measure how active a gene is.

genes from different organisms can be used to predict gene function in unknown, more complex genomes from simpler model genomes. This approach is known as *comparative genomics*. In *functional genomics*, the function of a gene is determined by inserting into bacterial cells which will make large amounts of the protein for which the gene codes, and then to determine the structure of the protein. The function can then be inferred in comparison to known protein structures and functions. *Functional genomics* disregards the gene sequence and the structure of the protein, and focuses instead on other properties of the protein product. The sources of data in functional genomics are summarised in table 1.1 on page 15.

In Section 3.1 we propose a formal concept that is suitable for a) the prediction of similarities in genome structure and function in species from which the sample sequence data is derived and b) modelling gene-interaction in gene networks. Apart from experimental genome analysis, for which results are already available in databases [35], the fuzzy relational factor-space approach will rely upon gene-expression data. These data are useful in complementing gene-coding-sequence-based *structural genomics*. As structural genomics – sequencing of entire genomes is progressing continuously, the focus is gradually shifting to *functional genomics*. Large-scale gene-expression assays are an important tool and formal methods for their analysis are yet to be developed.

Erwin Schrödinger, in his famous essay "What is life?", considered the question of how the events in *space* and *time* which take place within the spatial boundary of a living organism can be accounted for by physics and chemistry. Here I shall instead take a different route, trying to describe natural systems using mathematics. Considering that the models Schrödinger envisaged, were mathematical, it may seem like a contradiction. The idea is to describe biological principles ("natural laws") as they are known (measured, observed or perceived) by the biologist rather than describing what the mechanisms are in themselves (as described by biophysical and chemical models). Surely we cannot claim to be more accurate or even to have an alternative, since ultimatively the approach rests upon knowledge obtained through biochemical experiments (for which however the focus is on interactions rather than on investigations into the material structure).

1.2.1 Metabolic Engineering

Although the approach developed in this text is to investigate gene expression, gene interaction and gene function, these issues are not entirely unrelated to biochemical or metabolic engineering which is used to improve industrial organisms using modern genetic tools. The aim is to study physiological consequences of gene changes to allow inferences about the connections between genes and cell function.

Biochemical modelling of metabolic systems[17] is usually employed to explain specific phenomena by restricting oneself to essential factors, that is, trying to describe a pathway by a model in which groups of reactions are combined into overall reactions using kinetic laws. Kinetic laws of biochemical reactions are based on the notions of *concentration* and *reaction rate*[18]. A model of a metabolic pathway consists then of nodes, where each node represents a metabolite in reaction and each link shows a reaction in the pathway. Links are labelled by the rate of reaction. In general, kinetic rate laws are nonlinear functions of metabolites and solutions to differential equations are found using numerical algorithms. The book by Heinrich and Schuster [21] provides a comprehensive discussion on deterministic kinetic modelling of biochemical reaction systems. For a review of and commentary about mathematical modelling in biochemical engineering and lessons from metabolic engineering for functional genomics and drug discovery see [2] and [3], respectively.

Metabolic engineering is about the analysis and modification of metabolic pathways [71] and can be based on *metabolic flux analysis* [72, 14]. *Flux Balance Analysis* (FBA) allows the quantitative interpretation of metabolic physiology based on experimental data. The mathematical formulation is based on the conservation of mass, expressed in mass balance equations which describe all relevant internal, in- and outgoing metabolite fluxes of the cell. The concise description in [72, 14] is restated here. A metabolic network with m metabolites x_i, their unknown amounts and concentrations represented by a m-dimensional vector $\mathbf{x}$ are changing according to the vector differential equation

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{S} \cdot \mathbf{v} - \mathbf{b} \ ,$$

where $\mathbf{v}$ is the vector of n metabolic (reaction) fluxes and $\mathbf{S}$ is the stoichiometric $m \times n$ matrix that contains information on the reaction stoichiometry of cellular metabolism. The rows and columns of the matrix are associated with the metabolic balances and the metabolic fluxes, respectively. Any particular element s_{ij} indicates the amount of the i^{th} compound produced per unit flux of the j^{th} reaction. $\mathbf{b}$ is the vector of known metabolic demands, it is the net transport of metabolites out of the system under consideration. The time constants of metabolic transients are small in comparison to cellular growth rates and the dynamic changes in the organism's environment. One can therefore simplify the model for steady-state behaviour :

$$\mathbf{S} \cdot \mathbf{v} = \mathbf{b} \ .$$

This system of equations is usually underdetermined (i.e., the number of fluxes exceeds the number of metabolites) leading to more than one solution. The

[17] A *metabolic system* is considered as a network of enzyme-catalysed reactions in the cell.
[18] The *concentration* is defined by the number of moles of a given substance per unit volume. The *reaction rate* is expressed as concentration change per unit time.

null-space[19] of the matrix **S** defines a solution space, which Palsson refers to as the *metabolic genotype*. It describes all allowable flux distributions (reactions) by a set of metabolic genes. Any particular solution defined by a specific set of conditions is then called *metabolic phenotype*. The size of the null-space represents the flexibility of a cell to determine its metabolic capabilities. Metabolic flux distributions can be measured experimentally but only with limited accuracy.

According to Palsson, it is expected that flux balance analysis, in combination with experimental genomics, will play a role in establishing genotype-phenotype relationships. For this to happen, models that relate to gene expression or gene regulation are required. A major problem with metabolic engineering is that, in order to restrict the complexity of the model, a common assumption is that a single gene's product has a significant effect on the biochemical networks that are considered. However, there are growing doubts that such simplification can be sustained for long [3]. The developments in metabolic engineering, based on systems of differential equations modelling biochemical dynamics, are expected to find continued interest for industrial biotechology but with regard to functional genomics will be limited to very specific aspects of molecular systems. Instead of modelling metabolic pathways and regulatory networks as *mechanisms*[20], in the realm of Newton's mechanics, subsequent sections will elaborate on the *informational networks* that operate cells.

1.2.2 Gene-Expression Data

In DNA- or micro-arrays each probe cell in the chip contains a large number of different single-stranded DNA pieces of the region of a gene that codes for RNA (which in turn is translated into protein). To determine gene activity, a flourescently labelled single-stranded sample RNA is added, and if correctly matched, the flourescent RNA sticks to the complementary strand on the chip and emits a light signal. Relatively new techniques such as serial analysis of gene expression (SAGE), high resolution 2D gel electrophoresis and microarrays (gene or DNA chips) are a means to identify gene products and their quantity [85].

With array technology, it is now possible to study expression patterns within a variety of gene families or to search for new homologous genes. DNA arrays can be used in a variety of ways, which may be classified as genotyping

[19]The null space of a matrix is also called the *kernel* of the matrix. Generally, the concept is applied to any linear function (such as the distance) between vector spaces $V \rightarrow V'$ and the kernel of **S** is defined to be the set of elements of V which map to the zero vector in V'.
[20]We return to a discussion and critique of this approach to describe biological processes in Section 3.4.

and gene expression. Genotyping arrays are designed to examine DNA at the sequence level. Known sequence variants of a gene or collections of genes are represented in an array. A gene of unknown sequence can then be rapidly screened for a large number of changes. For example, in comparative genomic hybridisation one can correlate gene expression with disease states. DNA arrays for examining gene expression, also called 'gene chips', can involve longer fragments of synthetic or complementary DNA. The objective is to analyse gene expression levels to describe gene interactions in metabolic/regulatory pathways and hence to suggest unknown gene function. Expression data are the basis for gene-networks introduced in the next section. Mathematical models of gene interactions identified from expression data are required to have exceptional generalisation properties and are required to cope with considerable levels of uncertainty arising from fluctuations in the light source, fluorescence scattered from adjacent samples in the array, and a host of other experimental factors.

In addition to sequence information which allows us to relate genes via sequence similarity, gene-expression information can be used in defining gene relationships [17]. The data from gene microarrays can be sampled over time and pattern recognition techniques (e.g. clustering, principal component analysis) are used to identify genes with related functions. A recent discussion and overview for the use of gene-expression microarray data is given in [85]. It is expected that microarray data will play an increasingly important role in functional genomics[21]. As microarray facilities becomes available in many institutions, the area is progressing very fast. However, to this date, experiments are complicated, expensive and time consuming which is the reason why experiments are seldom repeated and in the case of time-series experiments only few sampling points are considered. Subsequently data are often imprecise and unreliable. For the foreseeable future, if not in general, sets of numerical data obtained from measurements, will not provide sufficient information for decision making (classification, prediction) in the presence of uncertainty. Even the most sophisticated mathematical techniques or data mining tools still require substantial knowledge and understanding of the process under consideration. The dilemma is reflected in Nobert Wiener's complaint[22]

> *"I may remark parenthetically that the modern apparatus of the theory of small samples, once it goes beyond the determination of its own specially defined parameters and becomes a method for positive statistical inference in new cases, does not inspire me with any confidence unless it is applied by a statistician by whom the main elements of the dynamics of the situation are either explicitly known or implicitly felt."*

[21]See for example the web-site of the European Bioinformatics Institute at http://www.ebi.ac.uk/

[22]Nobert Wiener: *Cybernetics: Control and Communication in the Animal and the Machines*, 1961.

The IMDS project at the Technical University Delft [27] is another example which shows that in future, efficient, direct measurements on the molecular level will be possible. IMDS (Intelligent Molecular Diagnostic Systems) is a multi-disciplinary research project at the Delft University of Technology. The IMDS will consist of two basic components: a measurement device and an information processing unit (IPU). The measurement device is a chemical sensor on a chip, which will be capable of rapidly performing vast numbers of measurements simultaneously. The IPU transforms the complex, raw measurements (of concentrations or optical signals) obtained from the sensor into output that can be employed as high-level decision support in various application domains. The focus is on unravelling the metabolic processes and the associated regulatory mechanisms of yeast.

The factor-space approach developed in Section 3.1, is aimed at a better understanding of the mathematical description of the interaction of genes and their function. It will therefore require information and data from more than one source. Results of experimental genome analysis are stored in databases accessible via the Internet. The progress of such world-wide projects is monitored by for example the Genomes OnLine Database (GOLD) [35]. For the approach presented here we however need access to a single source of genomic information in a well defined and structured format. An example for such a single data source providing an effective description and management of genomic information is the GIMS project [53].

Table 1.1 The four 'oms' defining different levels of analysis, providing different types of information and using different experimental techniques [49].

Level of Analysis	Definition	Method of Analysis
GENOME	Complete set of genes of an organism	Systematic DNA sequencing
TRANSCRIPTOME	Set of messenger RNA molecules present in a cell, tissue or organ	Hybridisation arrays, SAGE, High-throughput Northern Analysis
PROTEOME	Set of protein molecules present in a cell tissue or organ	2D-gel electrophoresis, peptide mass fingerprinting, two hybrid analysis
METABOLOME	Set of metabolites (low-molecular weight intermediates in a cell, tissue or organ)	Infrared spectroscopy, mass spectroscopy, nuclear magnetic resonance spectrometry

An effective study of gene expression, gene regulation and interaction of genes in a network will be a multi-levelled approach incorporating information from all four stages of gene expression starting with DNA via RNA, to proteins

and metabolites. Corresponding to each of these components biologists have defined four levels of analysis[23] as listed in table 1.1.

1.2.3 Gene Networks

Though genes and their individual functions become known, the study of the interactions of several genes in cellular functions is not yet developed to a large extend. The coordinated function of multiple genes has been described as *genetic circuits* or *gene networks* [50, 70]. Such genetic system represents a cellular network of gene products that together make up a particular function. More specifically, the genes in a network co-regulate one another's expression rates where each gene encodes a protein that serves as a regulator for at least one other gene in the network, influencing the rate of transcription, translation, or post-translational modification. Figure 1.6 outlines the general idea of genetic circuits as described by Palsson [50]. In [85], an overview is given on how gene-expression data from microarrays can be used for inferring gene networks.

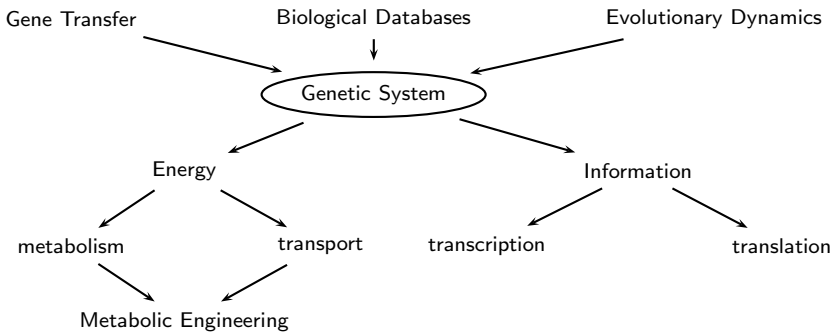


Fig. 1.6 Genetic systems, once expressed are autonomous. The aim of the fuzzy relational factor space approach is to eluminate the relationship between pathways and genetic systems. (Adapted from [50]).

The complexity of gene interactions is due to subprocesses, interacting on different levels and relationships which cannot be inferred simply by taking the system to pieces. In case of gene regulatory systems, we often observe

[23] Whereas *Genetics* describes the study of gene function in relation to the phenotype, the area of *Genomics* studies gene function related to sequence data. *Functional Genomics* is then closely related to the four 'oms' defined in table 1.1. *Proteomics* is understood as the study of protein structures and their prediction from sequence data as well as the identification of proteins from mass spectroscopy experiments. The *post-genome era* is expected to be primarily concerned with gene interactions, gene regulation.

processes in 'closed-loop' configuration. That is, measurements supposedly representing input-output relationships are in fact not from *a* model but a system that could be described as consisting of an *internal model* and a *controller*. The process under consideration is observed "under control". If we are to view biological regulatory processes as a control system, we are compelled to formulate Rosen's *anticipatory systems* in terms of *model-based predictive control*. The principle of self-reference is apparent in internal modelling providing a means for the system to both adapt and evolve. Although on the surface, internal modelling is akin to *adaptive control*, in anticipatory systems the control action is a function not only of current and past states but also of (model based) predictions of future states. Note that viewing gene regulatory processes as control systems is only a metaphor[24].

[24]In a metaphor we liken some process or phenomenon observed in one domain to a seemingly parallel process or phenomenon in a quite different domain.

2

A System-Theoretic Epistemology of Genomics

This chapter is of fundamental importance for all subsequent section. It is organised as follows. In the first part, there are six main propositions, preceded by a definition. Referring mainly to the philosophy of Arthur Schopenhauer[1], the first proposition provides the foundation for a phenomenological perspective of science and describes the certainty of uncertainty in any scientific enquiry; following the bad news, the second proposition recovers objective knowledge within the realm of the world of experience. The third proposition formalises the scientific method in form of Robert Rosen's modelling relation and introduces a system theory based on sets and relations. Proposition four describes entailment structures that are a basic tool of science. Proposition five describes Schopenhauer's 'differentiation' as the basic mode of operation for human minds and the last proposition draws some conclusions from differentiation. I briefly mention some personal conclusions on the consequences of the approach and provide examples for the conceptual framework presented. The examples are used to introduce the fuzzy relational model of gene-expression and function which we shall develop throughout the remaining text.

[1]The summary provided here follows closely the excellent discussion of Schopenhauer's philosophy by Brian Magee [38]. The main difference between Schopenhauer's and Kant's work is that Kant focussed on the nature of conceptual thinking while Schopenhauer focussed on the nature of experience.

2.1 PHENOMENAL CONSTRUCTIONS...

DEFINITION 1: The world of experience is Kants world of the ***phenomena*** - the empirical world (Wirklichkeit). A *phenomenon* is a collection of related percepts suggesting causal entailment.

PROPOSITION 1: If there is something that is grasped, then there is something that grasps it and everything that is said, is said by someone.

Proposition 1.1: The world as we ***experience*** it, is dependent on the nature of our apparatus for experience, with the consequence that things as they appear to us, are not the same as they are in themselves. Experience divides into two aspects: *perception* and *conception*.

Definition 1.1: ***Perception*** is tied to the phenomenal world - the world of cognisable objects (sensory impressions or percepts), which we observe and measure, and with which science deals. Perception is the process of *discerning* cognisable objects; to distinguish, to differentiate them. To *organise* percepts is a primary function of the ***mind***; it means to establish relations between them (cf. Definition 4.3). An example of perception is *understanding* (Verstand), the capacity for preconceptual, intuitive knowledge.

Definition 1.2: ***Conception*** is part of the world of concepts (ideas) in which we establish a *modelling relation* (cf. Proposition 3) between the self[2] (mind) and its ambience (the experienced, context, observed. Cf. Definition 3). Conception is the comprehension of phenomena. An example of conception is *reason* (Vernunft), the capacity to form and employ concepts based on the prior intuitive grasp of things.

Proposition 1.2: The world as we know it is our interpretation of the ***observable facts*** in the light of ***theories*** that we ourselves invent/construct. Within a theory, every argument has to have an absolute minimum of one premise and one rule of inference (e.g a relation representing IF A, THEN B) before it begins, and therefore begins to be an argument at all.

Proposition 1.3: Every argument has to rest on at least two undemonstrated assumptions, since no argument can establish either the truth of its own premise or the validity of the rules by which itself proceeds.

[2]Here we identify 'the self' with a human being's mind and intellect (understanding, reason), as opposed to his or her body. The self exists in a subject-object relation to its ambience (Proposition 1), describing the world as representation. Another aspect of the self, not discussed here, amounts to what Schopenhauer designates as *will*.

Proposition 1.4: Popper: Theories are formulated as to correspond in some useful way to the phenomenal world, whatever that may mean. The quest for precision is analogues to the quest for certainty and both – precision and certainty are impossible to attain.

Proposition 1.5: ***Uncertainty*** (the lack or absence of certainty) creates alternatives and hence choice. Wittgenstein: What we cannot speak about, we must remain silent about. What we cannot think, we cannot think, therefore we also cannot say what we cannot think.

DEFINITION 2: The world of things[3] (objects) as they are in themselves is Kant's ***noumena*** (Realität). Though we can have knowledge about the noumena, we can never have knowledge of it.

PROPOSITION 2: Kant, Schopenhauer: Reality is hidden but transcendentally real. The world of objects is representation, conditioned by the experiencing self (his mind), but has transcendental reality. The trancendental ideal (noumenon) and the empirical real (phenomenon) are complementary. Whatever is noumenal must be *undifferentiated*.

Proposition 2.1: Science deals with ***concepts*** to interpret aspects of the phenomenal world. Science does not describe an independent reality; it does not deal with the things what they are in themselves, but with phenomena through objects and relations defined among them. In other words, the aim of science, mathematics and philosophy is the study of *natural-* and *formal systems* (cf. Proposition 3).

Definition 2.1: An idea or concept is defined by

i) its ***extension*** - the aggregate of objects relevant to the concept.

ii) its ***intension*** - the collection of *factors* and their *attributes* describing it.

The two most important concepts by which our experience is made intelligible to us are ***space*** and ***time***, constructed to describe ***causal entailment*** (Definition 4) in the world of experience.

Definition 2.2: An ***object*** can be a physical (material) object or mass but also an abstract mathematical object or a concept. A multiple of objects defines a ***set***. An object is never the thing-in-itself, but something the cognising (perceiving and conceiving) self (mind) has constructed by discerning it from its context.

[3]We are using the word *'thing'* in a very general sense, so that it represents *anything* (e.g., objects, entities, qualities, properties, etc.).

Definition 2.3: We refer to a perceptible or cognisable quality of a natural system as a ***factor*** (or *observable*). A factor[4] is described by the mapping $f\colon U \to X$ from a set of objects U to ***factor-space*** X. While U denotes the ***hypothesis-space*** in which we define or infer statements about a phenomena in question, X is also referred to as the *observation-space* or *state-space* in which measurements or observations are represented. Only events in X are directly perceptible to us. A factor induces *relations* on the set of objects and between the set of objects and the set of states. Factors serve as the vehicle through which interactions between natural systems (e.g the sensory apparatus of the self and its ambience) occur, and which are subsequently responsible for perceptible changes arising from interactions (cf. Proposition 6).

Definition 2.4: ***Attributes*** establish the relationship between the phenomena considered and its context; they capture ***semantic information***. Attributes are represented by the mappings

i) $\widetilde{A}\colon U \to L$ from the set of objects U to a space L. This mapping is called the extension of a concept in U.

ii) $f(\widetilde{A})\colon X \to L$ from the set of states into L. This mapping is called the ***representation extension*** of a concept in X. For L being the unit interval $[0, 1]$, these two mappings are referred to as ***fuzzy sets***[5].

Proposition 2.2: Objective knowledge of causal entailment (cf. Definition 4), is attainable within the realm of the phenomenal world. What is given to us in direct experience are the representations of ***sense*** (through perception) and of ***thought*** (through conception). The world of experience cannot exist independently of experience. Experience is objective but what is denied, is the validity of inferences from what we experience to what we do not experience. Scientific knowledge is common sense knowledge made more critically self-aware and raised to a level of generality.

DEFINITION 3: A ***system*** is a set of objects and relations defined on them. Formally, we define a system by the pair (U, R) where U is a set of certain things, i.e., objects u, and R is a relation defined on U or the Cartesian product $U \times U$, in which case we have $R \subset U \times U$. The

[4]The notation $f\colon U \to X$ is read as "a mapping f from space U to X". An element of U, denoted $u \in U$, as an argument to f maps to the value $f(u)$ in X; denoted $u \mapsto f(u)$.

[5]For a 'non-fuzzy' or 'crisp' set A, the degree of membership $A(u)$ can only take two values, zero or one, denoted $A\colon U \to \{0, 1\}$. Varying degrees of membership between zero and one, $u \in [0, 1]$, can be used to model different kinds of uncertainty (ambiguity, fuzziness, vagueness,...) and should allow us to integrate qualitative, context-dependent knowledge into the otherwise quantitative model.

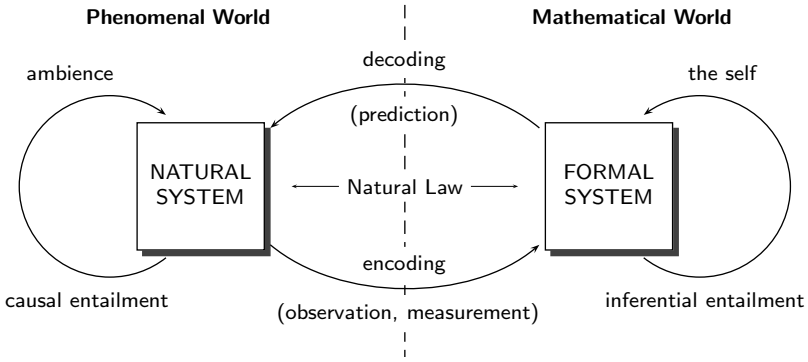


Fig. 2.1 The modelling relation between a natural system N and a formal system F [60]. If the modelling relation brings both systems into congruence by suitable modes of encoding and decoding, it describes a *natural law*. In this case F is a *model* of N, that is, N is a realisation of F.

relation(s) R role is usually to order, structure or partition elements in U. Systems do not exist independent of the mind but they are a formal representation of aspects of the phenomenal world. A ***formal system*** represents the interior world of ***the self*** while a ***natural system*** is an element of the outer or exterior world of ***the ambience*** (context), a set of phenomena in the world of experience. As such it embodies a mental construct (i.e., a relation established by the mind between percepts) serving as a hypothesis or model pertaining to the organisation of the phenomenal world.

PROPOSITION 3: Rosen: In order to understand (explain), one establishes a ***modelling relation*** between a natural system, N, and a formal system F. If the modelling relation brings both systems into congruence by suitable modes of ***encoding*** (measurement, observation) and ***decoding*** (prediction), it describes a ***natural law***. In this case, F is a ***model*** of N, or N is a ***realisation*** of F. Modelling, the process of establishing a modelling relation, bringing the two entailment structures into congruence, is a creative mental act, it is an ***art***.

Proposition 3.1: A model is the basis for ***reasoning***. Reasoning is the process of turning ***facts*** into knowledge. ***Knowledge*** is the result of understanding (explanation, experience) and is represented by law-like relations. A ***law*** (or *principle*) can only describe what a natural system is like, not what it is.

Definition 3.1: A fact is a context-independent measure extracted from ***data*** (e.g. measures of variability or central tendency). A descriptive or ***fact explanation*** (e.g pattern) is the use

of a theory and data to induce a singular factual statement. A ***law-like explanation*** (e.g rules) uses a theory, subsidiary assumptions (statements, axioms) and data to infer a law.

Definition 3.2: Data are instances of states, i.e., evaluations of objects using factors. Data are context-dependent as is knowledge. The process of collecting data is referred to as ***measurement***. The estimation of parameters of a formal model from data, is referred to as ***system identification***.

DEFINITION 4: By separating the observed aspect of the phenomenal world from the formal model and the self observing it, the following two kinds of objects and entailment are fundamental:

i) Objects in natural systems are referred to as ***components***. The realisation of relations in a natural system is referred to as ***causal entailment*** (causality).

ii) Objects in formal systems are referred to as ***propositions***. The evaluation of relations in formal systems is referred to as ***formal entailment*** (inference).

PROPOSITION 4: To ask "why u?" is to ask "what entails u?". To understand entailment is the sole function of the understanding and its only power. Conversely, all entailment and consequently the whole of reality, is only for the understanding, through the understanding, in the understanding. Understanding, through inference, is the subjective correlate of causal entailment.

Proposition 4.1: Entailment exist only between objects in the phenomenal world. The succession of events or phenomena is not arbitrary; there are relations manifest in the world of phenomena and these relations, at least in part, can be grasped by the human mind.

Definition 4.1: The concept of ***linkage*** between factors represents causal entailment in natural systems. The linkage between any two factors is a relation determined by comparison of the partitioning (equivalence relations) induced by the two factors.

Definition 4.2: For a factor $f : U \to X$, in a formal system, object $u \in U$ entails $f(u)$. Asking "why $f(u)$?" is answered "because u" and "because f". The former corresponds to Aristotle's ***material cause*** of 'effect' $f(u)$, while the latter refers to the ***efficient cause*** of $f(u)$.

Proposition 4.2: For entailment to exist, an act of ***differentiation*** is required. Each time we refer to anything (whether a percept or concept), we are specifying criteria of distinction, ***discerning*** an object from its ***context***.

Proposition 4.3: Discerning an object, we implicitly recognise organisation.

Definition 4.3: ***Organisation*** is defined by ***relations*** that must be in place in order for something to exist (to be there, to be an object). If a system exhibits a particular behaviour, it must possess certain properties producing the behaviour. These properties will be called the organisation of the system. If the organisation of a system does not change, the organisation is also referred to as the ***structure***.

Definition 4.4: ***System theory*** is the study of organisation *per se*. It defines formal systems by means of mathematical relations (equality, elementhood, subsethood, greater than, smaller than, ...) and set ***comparisons*** (union, intersection, and complement).

Proposition 4.4: For anything to be different from anything else, objects, sets and concepts have to be presupposed.

Proposition 4.5: Causality manifests itself only through changes in states, called ***state-transitions***, leading to sequences of states, entailing an effect that is again a state. The change of a particular state is called an ***event***.

DEFINITION 5: Anything that is observed is subject to ***change*** as for anything that was there, it has changed (is different) through differentiation.

PROPOSITION 5: Discerning is an ***interaction*** that brings forth an object. Knowing is doing (discerning); doing is understanding (experiencing). Knowledge arises from the plurality and separate existence of beings (objects); knowledge arises from and through ***individuation*** (differentiation).

Proposition 5.1: Discerning, implies change, *reveals* diversity and complexity but also *imposes* order.

Definition 5.1: A particular time-invariant relation, specified for a set of quantities and a resolution level, and based on samples of a certain pattern, will be called the ***behaviour*** of the corresponding system. If the behaviour of the system can change, the behaviour is also referred to as ***dynamics***.

Proposition 5.2: Although differences may exist (through differentiation), knowledge of it and of uncertainty leaves a choice to the nature of entailment.

Proposition 5.3: Although knowledge originates *with* experience, it does not all arise *out of* experience. Apart from understanding

through observation or contemplation alone, the observation of change through manipulation is a means to gaining knowledge.

Definition 5.3: Creating a new perturbated system which can be compared with the original, the discrepancy between ***behaviours*** determines its ***function*** while discrepancies between system ***structures*** determine its ***components***.

Proposition 5.4: There is no such thing as knowledge of knowing since this would require that the self separated itself from knowing and yet knew that knowing.

DEFINITION 6: Learning is the process of gaining knowledge through experience (perception and conception). There are two modes of pursuing knowledge: ***contemplation*** and ***manipulation***.

PROPOSITION 6: Living is learning; learning is experiencing; experiencing is discerning; discerning is an (inter)action; an interaction brings forth a change (difference). The interaction between a natural system and our sensory apparatus generates percepts from a change or modification within it. The sensory apparatus itself is a natural system, and we can say that the interaction of any two natural systems causes some change which we can represent by means of factors. Changes make the world comprehensible.

Proposition 6.1: Differentiation is the essence of life, as we perceive and decide it.

Proposition 6.2: The pursuit of knowledge provides a choice between contemplation and manipulation.

Proposition 6.3: ***Tolerance*** is the appreciation of diversity through contemplation. ***Morality*** derives from the knowledge that, since the noumena is undifferentiated, differences are only transcendentally real.

2.2 DISCUSSION

The previous section outlined the basis for a system-theoretic epistemology integrating aspects of Arthur Schopenhauer's philosophy, Robert Rosen's system theory and Peizhuang Wang's factor-space theory [73, 37]. With the work of Immanuel Kant, metaphysics was discovered in the subject. Kant identified the concepts of space, time and causality as *a priori* and therefore conditional for experience[6]. He also showed that these apply only to experience and may

[6]In his famous essay "What is life?" [66], the physicist Erwin Schrödinger, comes to the conclusions that "our sense perceptions constitute our sole knowledge about things. This objective world remains a hypothesis, however natural".

not be used to found a metaphysical system. Our mind organises the elements of experience to the principle of causality, but in contrast to Davide Hume, who derived causality *from* experience, Kant showed that we approach the world around us *with* the principle of causality already being there. With Kant, the subject therefore becomes central to reasoning and understanding. The subject guarantees the unity of the outer world, the knowledge of my being is the basis for the re-presentation of the world we experience. In the words of Werner Heisenberg, "What we can observe is not nature itself, but nature exposed to our method of questioning." With the creation of a domain in which pure reason allows for certainty and truth, we also create the noumena as something which is forever unaccessible. Kants 'things as they are in themselves', the *noumena*, we ourselves create by the knowledge of the *phenomena*. While others, namely Fichte, Schelling, Hegel and Marx, tried to fill the gap of uncertainty created by Kant, Schopenhauer accepted the presented limitations, refined the boundaries and clarified our knowledge about the noumena. For everything that becomes part of our experience, we are 'forced' to ask for causes and entailment.

According the type of objects we deal with, Schopenhauer describes in his dissertation *'The Fourfold Root of Sufficient Reason'* the different ways by which we establish such entailment relations. According the type of objects we deal with, Schopenhauer describes in his dissertation *'On the Fourfold Root of Sufficient Reason'* [65] the different ways by which we establish such entailment relations. Schopenhauer asserts that the everyday world is made up of objects of four classes; the first class consisting of material objects, such as the chromosomes in the genome; the second class consisting of concepts and combinations of concepts, such as gene function or hypotheses regarding gene expression; the third class consisting of time and space; and the fourth class consisting of particular human wills. These objects are interconnected in a number of ways, allowing questions to be asked and answered; there is always a reason. Material objects are subject to change, and of any change the question "Why does it occur?" can be asked. Concepts combined in appropriate ways constitute judgements or statements which can be questioned by asking "Why is it true?". Third, time and space are represented by mathematical objects for which we can ask "Why does it possess its characteristic properties?". Again, there is always a reason - a 'sufficient reason'. The four forms of *the principle of sufficient reason* are that every change in a material object has a cause; the truth of every true judgement rests upon something other than itself (cf. Proposition 1.2 and 1.3); all mathematical properties are grounded in other mathematical properties; every action has a motive. Objects of the four classes comprise therefore those, being subject to change (first class), those bearing truth (second class), those possessing mathematical properties (third class), and those of the fourth class giving rise to actions under the influence of motives. In science, formal systems are used to model natural systems; to establish concepts; to describe relations between percepts;

and to make predictions. In science, formal systems are used to model natural systems; to establish concepts; to describe relations between percepts; and to make predictions. Science is the description (comprehension) of the phenomenal world. The 'natural sciences', physics, chemistry and biology are based on *comparisons* (using sets – union, intersection, and complement) for the purpose of *reasoning* (*classification* based on *transitive laws*). Mathematics is concerned with the construction of formal systems using *abstract sets* and *formal relations*. Philosophy studies the consequences and foundations of science and mathematics. Relating natural systems with formal ones, we aim to make inferences in the latter to make predictions about the former.

Ultimate or philosophical explanations are not to be looked for in science[7] (Proposition 1.2)[8] because the applicability of science is confined to the phenomenal world (Proposition 2.1). Our experience is made intelligible to us in terms of space, time, and causality; for only then it is possible to talk of there being more than one anything, or of anything being different from anything else. Differentiation, discerning and individuation are at the root of experience and therefore science. The possibility of plurality (Schopenhauers *principium individuationis*) is necessarily conditioned by time and space. If the mathematical structures we employ to encode natural systems, are not in themselves the reality of the natural world, they are the only key we possess to that reality. The essence of the modelling relation (Figure 2.1) is that we have to explain the correspondence between natural systems and its mathematical representation. There are many examples of the remarkable correspondence between mathematical models and the behaviour of the natural world, but is must be admitted that no one of these is final. The modelling relation, here used as a conceptual device to clarify the relationship between natural systems and mathematical structures created for understanding such systems, is in fact a model of the scientific method; providing an intriguing subject for further study and contemplation. (See for example [62]).

Knowledge is, of its nature, dualistic: there is something that is grasped and something else that grasps it. The whole world of objects is representation, conditioned by the *subject* (the self or observer, an object himself); it has ***transcendental reality*** (Proposition 1 and 2). All knowledge takes the subject-object form, but only in the world of phenomena can subject and

[7]Or as Henri Poincaré suggested, the aim of science is not things in themselves but the relations between things; outside these relations there is no reality knowable. Schopenhauer's 'principle of sufficient reason' explains connections and combinations of phenomena, not the phenomena themselves.

[8]In the words of Ludwig Wittgenstein (Tractatus Logico-Philosophicus): "The sense of the world must lie outside the world... What we cannot speak about we must remain silent about... What can be described can happen too, and what is excluded by the laws of causality cannot be described."

object be differentiated (Definition 1.1). According to Schopenhauer, and in contrast to Kant, the world we perceive is not just indirectly constructed by conception (Definition 1.2) and concepts (Proposition 2.1) we use to describe them but already directly by the sensory apparatus. Perception (Definition 1.1) is intellectual in the sense that objects are created by the intellect; it is not a matter of bare sensations. According to Schopenhauer the world of perceptible objects is the creation of the faculties of sensibility and understanding. Our intellect is presented with sensations or sensory data, upon which it imposes the concepts of time, space and causality. We could say that perception (Definition 1.1) provides the letters or words, by which the mind forms the words and sentences, respectively. Although independent reality is something which human knowledge can approach only asymptotically, never to grasp or make direct and immediate contact with, there exists objective knowledge in the realm of the phenomenal world. We may not describe the things as they are in themselves, the objects however have empirical reality. Kant's transcendental idealism ensures empirical realism, while ignorance to the distinction between the things in themselves and the appearances (transcendental realism) results in scepticism about the knowability of objects (empirical idealism). A common error is to mistake the gap between the phenomena and noumena with a lack of objective knowledge in the phenomenal world or to fill the apparent gap between the phenomena and noumena with some form of *subjectivism*, *relativism*, pessimism or religious belief instead of asking further questions. Following Poppers 'critical rationalism', we ought to combine an empiricists view of reality (empiricists ontology) with a rationalist view of knowledge (rationalist epistemology).

The scientific method, relying on the concepts of space and time, investigates objects (whether physical or abstract) and establishes relations between them (Proposition 3). In order to understand or know a natural law (principle), i.e., to establish the existence of the modelling relation (Figure 2.1) between a natural and formal system, two further concepts *regularity* and *repeatability* play an essential role. Regularity is associated with the existence of *relations* while repeatability is the basis of *comparisons*. In simple terms, we may require the repetition of an experiment in order to establish regularity through comparison. To *decide* upon regularity or *chance*, we need repeatability; Chance and *randomness* are defined by *irregularities* – the absence of relations. See also Figure 2.2.

The notion of *existence* causes further problems as one may ask whether we mean "does not exist *in principle*" or whether we mean "is not accessible, observable, not knowable" without refined means of observation or measurement. A chance mechanism induces randomness, a form of uncertainty which makes certain events or states *unpredictable*. Whether with refined measurements and tools, by "zooming in", we could identify such relations, say on a "microscopic" level, introduces the notion of *scale* or *scaleability*. To al-

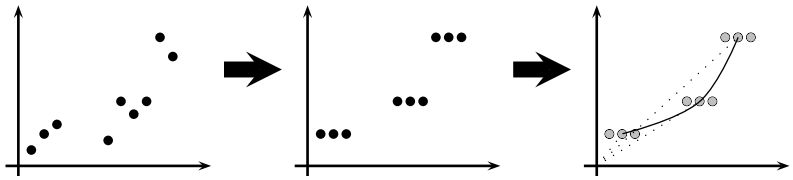


Fig. 2.2 Repeatability, comparison, modelling and the uncertainty in fitting a model to data.

low reasoning in the presence of uncertainty, we may accept the notion of randomness or chance as "undetermined through observation" and therefore view as if the process *is* by chance. Regarding Proposition 1.1 the question of whether an ideal organism, with perfect sensory apparatus, could know the noumena is irrelevant because it does not exist as an object of the phenomena. *Complexity* is commonly associated with the inability to discriminate the fundamental constituents of the system or to describe their interrelations in a concise way. Complexity is a characteristic feature of the (empirical) real. Nature's complexity is literally inexhaustible - as a result of the inherent limitedness of our knowledge of nature. Complexity also induces uncertainty. If the formal system is comparatively simple in comparison to the natural system it is to model, predictions from the model will become unreliable, we are forced to attach a measure of confidence (such as a probability) to predictions. Similarly, observations and cognition of complex systems will necessarily be vague, fuzzy or ambiguous if we try to identify interrelations and interactions. Like randomness, we therefore take the concept of complexity as closely related to that of *understanding*, to express uncertainty in understanding and reasoning rather than as a property of the system or data themselves. From our definitions and propositions above, understanding implies the existence of an *object-subject relation*, i.e., we assume the presence of a subject having the task of studying a natural system (objects, relations), usually by means of model predictions. Complexity is therefore related to both, the subject and the objects. The success of modern science is the success of the experimental method. The aim of modelling, whether using formal mathematical models or for instance the biologists expert knowledge and intuition, is to infer a natural law or fundamental principle which should yield non-ambiguous predictions. Whenever substantial disagreement is found between theory and experiment, this attributed either to side-effects of the measurement process or to incomplete knowledge of the state of the system. In the latter case, using a reductionist approach, we would seek to refine our measurements, i.e., improving accuracy or adding variables (factors) to measure.

The concepts of space, time and causal entailment in science are formalised by mathematical objects such as sets, order and equivalence relations (cf. Definition 2.1 and 2.2). If we denote an object by u, we write $u \in U$ to state that

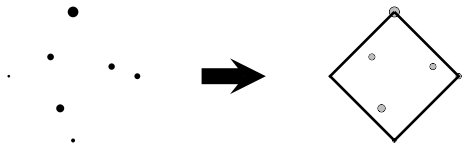


Fig. 2.3 Scaling in modelling.

u is an element of the set U. Before objects can be thought, a set in which these objects can be elements of must exist, not necessarily as an object itself but as a concept. If a set is empty, what remains is an empty set, denoted $\emptyset$. In order to apply a mathematical set, say for example $U = \{5, 3, 1, 2, 4\}$, in a real-world context, the set is usually furnished with an ordering relation because only then we are able to make *comparisons* in reference to U. Then $U = \{1, 2, 3, 4, 5\}$, as an ordered sequence, may be used to count for example events. On the other hands the comparison itself can structure the elements in U into *equivalence classes*, e.g. $\{2, 4\}$ and $\{1, 3, 5\}$, where elements share properties, are equivalent in a defined sense and would therefore not be distinguishable in measurement or observation. The set, endowed with a relation, or relations, defines a system. Representing a natural system by means of a formal system (cf. Definition 3), we encode it using factors f which map an object u into a point in the observation or factor space X. We here use the term space to denote the fact that X should be endowed with some (mathematical) structure allowing us to compare and order its elements, for example to define distances between points in X; leading to what is called a topological space.

Since sets of objects and relations play a central role in modelling natural systems, we should have a closer look at their definition. A *set* U is a collection of objects, called the *elements* of U. If u is an element of U, we write $u \in U$ and denote the set by $U = \{u\}$. Suppose two elements, first $u_1 \in U_1$, followed by $u_2 \in U_2$, are chosen; then this choice denoted by the pair (u_1, u_2), is called an *ordered pair*. The set of all such ordered pairs is called the *Cartesian product* of U_1 and U_2,

$$U_1 \times U_2 = \left\{(u_1, u_2) \text{ for which } u_1 \in U_1, \text{ and } u_2 \in U_2\right\} .$$

If furnished with some mathematical structure, a set is also referred to as a *space*. Any subset R of $U_1 \times U_2$ defines a relation between the elements of U_1 and the elements of U_2. A *relation* is therefore a set of ordered pairs, denoted

$$R = \{(u_1, u_2) \in U_1 \times U_2 \text{ for which } R(u_1, u_2) \text{ holds true}\} .$$

Since by R an element in U_1 is associated with one or more elements in U_2, R establishes a *multi-valued* correspondence :

$$\begin{aligned} R: \quad U_1 \times U_2 &\rightarrow \{0,1\} \\ (u_1, u_2) &\mapsto R(u_1, u_2) \ . \end{aligned}$$

An important family of relations are ***equivalence relations***, denoted $E(\cdot,\cdot)$. The equality relation, $=$, is an example. Equivalence relations are required to be *reflexive*, $E(u,u)$ holds for all $u \in U$ and *symmetric* $E(u,u')$ implies that $E(u',u)$ holds equally true for all $u, u' \in U$. The most important property of equivalence relations however is ***transitivity***: **if** $E(u_1,u_2)$ holds, **and** $E(u_2,u_3)$ holds, **then** $E(u_1,u_3)$ holds true as well. If u_1 equals or is similar to u_2 and u_2 equals or is similar to u_3, then u_1 also equals or is similar to u_3. Transitivity therefore provides a basic mechanism for reasoning; given two pieces of information (about u_1 and u_2, as well as u_2 and u_3) we can infer a third relation (between u_1 and u_3). If E is an equivalence relation on a set U, and if $u' \in U$ is any element of U, then we can form a subset of U defined

$$[u']_E = \left\{u \ : \ E(u',u) \text{ holds}\right\} \ .$$

Where the symbol ': ' is a short form of "for which" and if $E(u',u)$ holds true, we write $E(u',u) = 1$ and $E(u',u) = 0$ if it doesn't. The set $[u']_E$ is called ***equivalence class***. In figures 2.4 and 2.5 the areas described by factors are equivalence classes, representing sets of objects that have identical properties or which are not discernable by factor f. The set of equivalence classes of U under an equivalence relation E is called ***quotient set*** of U, denoted U/E. Considering any two ways of encoding a system, or alternatively changing (exciting, pertubating) one system to make two observations, we use the factors f and g to describe the modes of encoding/observation, the study of the *linkage* between the two factors f and g provides a basis for *reasoning*, i.e., will allows us to infer or validate entailment relations in the natural system under consideration. These ideas will be further elaborated in sections 3.5.2 to 3.5.4 when we devise an interface between our more abstract formal model and experimental data.

The present section described how we experience and learn (understand, gather knowledge etc). The basic principle of experience and therefore any scientific investigation is differentiation (cf. Proposition 4.2 and 5). All there is, is that which the subject brings forth in his or her distinctions. We do not distinguish what is, but what we distinguish is. We may say that the process of discerning therefore also creates or identifies diversity and alternatives; hence creating a choice, a choice to act upon the knowledge or experience. It is this point at which human behaviour defines the meaning of tolerance and morality (Proposition 6.3). Although recognition of diversity for some implies an appreciation of it, this is unfortunately not the case for a large proportion of the human species who take the principle of experience as the basis

for separating and discriminating against other species. There are two ways in which we can act upon diversity, to appreciate it or to use it in a way which, in the worst case, may lead to racism, capitalism and speciesism[9]. We may refer to these two ways to respond as 'contemplation' and 'manipulation' (Definition 6). Charles Darwin and Albert Einstein are probably the best examples of how observation and contemplation alone can create knowledge. In molecular biology, as in engineering, the design of experiments in which we manipulate, i.e., pertubate or change a system to study its properties is a central task (cf. Proposition 5.3 and Definition 5.3). As described in Proposition 6, change through interaction is a 'natural' aspect of experience and learning, which should not, cannot be restricted. The link to human behaviour and ethics only arises if we consider the *use* of the knowledge we gained. A similar perspective on man's action guided by illusionary perception, which is shaped by fragmentary thought was given by the physicist David Bohm [6]. His conclusions regarding theories as every-changing forms of insight and not descriptions of reality as it is, not only resonate with Popper's philosophy but also with Schopenhauer. In Bohm's view both, "relativity theory and quantum theory agree, in that they both imply the need to look on the world as an *undivided whole*, in which all parts of the universe, including the observer and his instruments, merge and unite in one totality". The physicist Erwin Schrödinger, who read Schopenhauer, refers to the principle of differentiation by discussing the apparent multiple of egos in (Western) thought ([66], *Mind and Matter*). He describes the reason why our sentient, percipient and thinking ego is met nowhere within our scientific world picture - "because it is itself that world picture. It is identical with the whole and therefore cannot be contained in it as a part of it. [The minds] multiplicity is only apparent, in truth there is only one mind". In his discussion of Kant's philosophy, Schrödinger acknowledges Schopenhauer's work. In contrast to Ludwig Boltzmann, who disliked Schopenhauer for his science, Schrödinger separates Kant's and Schopenhauer's philosophy from their attempts to find evidence for it in the sciences of their days. Schopenhauer is often misunderstood and his influence frequently ignored. A long list of scientists and philosophers acknowledges the philosophical tradition from Parmenides, Plato to Kant and Schopenhauer. Nietzsche and Freud are the most prominent representative who considered the 'human aspects', while Karl Popper specialised in the consequences for the sciences. We mentioned Bohm and Schrödinger but also Albert Einstein "has not - as you sometimes hear - given the lie to Kant's

[9]The effect of the principle of experience on society is well demonstrated by the use and meaning of the words 'discrimination' and 'exploitation'. Discriminating is making a distinction, to differentiate – a fundamental principle of life as described above, but also synonymous for a lack of appreciation of diversity. In fact, discrimination is a form of intolerance towards other beings. Likewise the word 'exploitation' comes from Latin *explicare* or 'explicate' – to make clear. Common use of the word is however to describe intolerance, say towards the environment.

deep thoughts on the idealisation of space and time; he has, on the contrary, made a large step towards its accomplishment." ([66], *Mind and Matter*, p.149). I was in fact Kant who made clear that there is no doubt that all our knowledge begins with experience but although our knowledge originates with experience, it dos not all arise out of experience.

This section outlined a system theoretic epistemology in the spirit of Arthur Schopenhauer. According to Schopenhauer we do what we want but we do it necessarily. This may lead to a rather pessimistic conclusion on the consequences of the described principles by which we operate, observe and manipulate the world around us. I hope to show that through the understanding, of the understanding we may have a choice, for the denial of Schopenhauer's *will*. He himself hinted at the possibility of a disposal of wants by grasping the illusory nature of the phenomenal world, and hence its nothingness, in order to gain some appreciation of the nature of the noumenal.

With regard to philosophy, we developed a 'constructivist' system science perspective based on Schopenhauer's philosophy but allowing for an 'existentialist' outlook on (human) behaviour. For the system theory, born out of the philosophical framework, the objective is to find a representation of molecular systems which is general and quite independent of their physical or chemical constitution. Such fuzzy relational biology is further motivated by the following examples. The first example is to illustrate the role of factors in perception and conception, the second example introduces Newtonian mechanics as the root of what has become the paradigm of mechanisms in general. The success of these models in some areas of science and technology has also led to their application in biotechnological processes (Example II). However, a further extension of these ideas to molecular systems and gene interactions has not been successful. Although bioinformaticians use descriptive statistics to extract pattern from data, formal mathematical models have so far played no role in the creation of biological knowledge in modern molecular biology. Example IV therefore suggests a phenomenological model which follows directly from the considerations in Section 2.1.

2.3 EXAMPLE I : "LEARNING IS DISCERNING"

In Figure 2.4, on the left, a space is depicted for which the objects are not discerned. Dividing the space as shown in the diagram on the right hand side, observing its objects, implies discerning those objects on the left from those on the right.

Instead of a vertical line we may have observed the objects in a different way, introducing a different factor (Figure 2.5, on the left). In mathematical terms, the mind imposes an *equivalence relation* that holds true for all

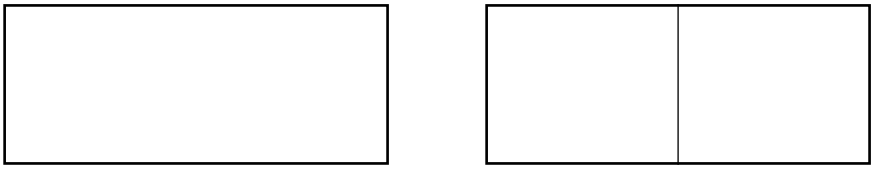


Fig. 2.4 In the space depicted on the left, the objects are not discerned while on the right the observation by means of some factor introduces a change, discerning objects on the left from those on the right.

elements indistinguishable within an *equivalence class*. We can then discuss the difference between the two modes of observation, i.e., the *linkage* between factors. The linkage between or comparison of factors therefore provides us with a means of reasoning and learning about the system (the set of objects relations defined upon them). We should however note that the explanation of the observation process itself required discerning. By drawing the box on the left in Figure 2.4 we had to discern the objects within it from those outside it.

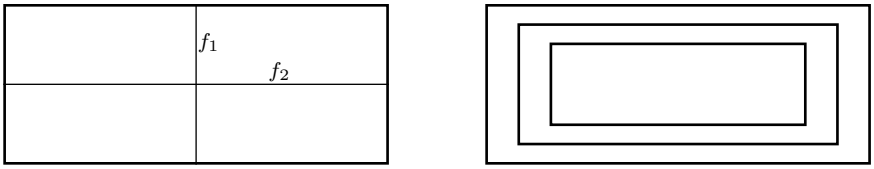


Fig. 2.5 Left: A change to the system will change the observation through the factor or equivalently, different means of observing by different factors provide distinct observations. Both ways, we can reason about the system by means of factors and the equivalence relations induced. Right: The explanation itself requires us to discern the objects within the box from those outside, leading to an infinite regress if we are to discuss 'the part' and 'the whole'.

2.4 EXAMPLE II : DYNAMICAL SYSTEMS

The arguments leading to and following Proposition 3 described modelling as a central part of learning through experience. As humans, so do other organisms use models (as an abstraction) for explanation or prediction. Organisms in general are therefore able to change their present behaviour in accordance with the model's prediction; the behaviour of biological systems is ***anticipatory***. As pointed out by Rosen, a formal system using a model based on differential equations only, is not able to describe such anticipatory or model-predictive

behaviour. Using systems of differential equations, the rate of change of a factor at any instant is expressed as a function of the values of other factors but cannot depend upon future states. Such systems are ***reactive***. Modelling dynamic systems with differential equations can often be expressed by a set of first-order equations :

$$\frac{\mathrm{d}f_j}{\mathrm{d}t} = \phi_j\big(f_1, \ldots, f_r\big) \ , \qquad j = 1, \ldots, r \tag{2.1}$$

where the rate of change of factor (observable, state-variable) f_j depends *only* on the present state defined by factors f_j. A simple example for (2.1) is a physical object u with mass m moving along a line under the action of a constant force denoted by F. Using Newton's law,

$$F = m \cdot \frac{\mathrm{d}v}{\mathrm{d}t} \quad \text{and} \quad v = \frac{\mathrm{d}x}{\mathrm{d}t} \ ,$$

where x denotes the displacement and v the velocity of the mass. For a particular system, a formal model can be defined by

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= v \\ \frac{\mathrm{d}v}{\mathrm{d}t} &= -\frac{\theta}{m} \cdot x \end{aligned}$$

where θ denotes a parameter specific to the natural system under consideration. Here the formal system uses two ***state-variables*** (factors) denoted by x and v, $f_1 \doteq x$ and $f_2 \doteq v$. The manifold of all possible states of the system, referred to as the ***state-space*** is illustrated in Figure 2.6. The *physical principle* described here is a conditional statement of the form

IF mass=m, force=F, THEN position=x and velocity=v.

Conceptual 'closure' of the system amounts to the assumption of constancy of the externally imposed force F. The model is deterministic in that the object's state at time t is fully determined from the *initial conditions* (of it position and velocity) and therefore permitting prediction of future states by integrating the set of differential equations. Newton's laws of motion, which state that the acceleration of an object is directly proportional to the force acting on it and inversely proportional to its mass, imply that the future behaviour of a system of bodies is determined completely and precisely for all time in terms of the initial positions and velocities of all the bodies at a given instant of time, and of the forces acting on the bodies. These forces may be *external forces*, which arise outside the system investigated, or they may be *internal forces* of interactions between the various bodies that make up the system in question. Although modelling in the Newtonian realm has proved successful in a number of engineering applications, the representation

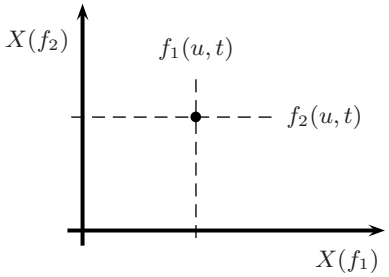


Fig. 2.6 The state-space (factor-space) of a simple dynamical system modelled by two state-variables (factors).

of objects and their behaviour in form of differential equations fails to extend to a large number of objects.

A *solution* to equation (2.1) is an explicit expression of each factor f_i as a function of time, $f_i(\cdot, t)$. A particular solution, defined by initial conditions, corresponds to a curve in the state-space, called ***trajectory***, and describes the evolution of the system over time. The passage of time implies the concepts "before" and "after", stated formally by the ***transitive law*** $t_1 < t_2$ and $t_2 < t_3$ implies $t_1 < t_3$ for the binary relation $<$.

Note that differential equations may be used to model a specific form of causal entailment in natural systems, the equations by themselves however do not state that changes are *produced* by anything, but only that they are either *accompanied* or *followed* by certain other changes. Considering $\mathrm{d}f/\mathrm{d}t = \phi(t)$ or equivalently $\mathrm{d}f = \phi(t) \cdot \mathrm{d}t$, it merely asserts that the change $\mathrm{d}f$ undergone during the time interval $\mathrm{d}t$ equals $\phi(t) \cdot \mathrm{d}t$. The notion of causality is not a syntactic problem but a semantic one; it has to do with the interpretation rather than with the formulation of theories or formal systems.

2.5 EXAMPLE III : METABOLIC SYSTEMS

The reactive paradigm of dynamic systems models using differential equations, described in the previous example, has also been applied to biotechnological processes and systems of genes interacting. For (autocatalytic) biochemical reactions of an (aerobic) biological process a substrate S is turned into a biomass x by consuming oxygen O. The process is characterised by the specific biomass growth rate, depending on the consumption rates of the substrate and oxygen:

$$S + O \longrightarrow P \ .$$

The *biochemical principle* described here takes the form of a conditional statement

$$\text{IF substrate}=S, \text{ oxygen}=O, \text{ THEN biomass}=X.$$

With three state-variables f_1-substrate concentration, f_2-biomass concentration and f_3-oxygen concentration we can define a set of differential equations in the form of equation (2.1). These equations are usually non-linear, and the inability to solve them forces us to make various assumptions and simplifications. For specific biotechnological processes, investigated in metabolic engineering, these assumptions are often valid but nevertheless limit our ability to understand more complex systems of gene interactions investigated in the field of genomics.

Gene interactions can be represented by their effect on the synthesis rate of gene products. Studying gene interactions or gene-networks, concentrations of gene products are therefore chosen as the state-variables. The change of concentrations of proteins over time (the left part of equation 2.1) is governed by direct regulation of protein synthesis from a given gene by the gene products of other genes (including autoregulation as a special case); transport of molecules between cell nuclei; and decay of protein concentrations.

The problem is that perturbations to cells have multi-gene, multi-transcript, multi-protein response but for the theory to remain tractable, one usually has to assume a single gene's product having a significant effect on the biochemical network. The reductionist strategy to analyse more complex systems has therefore been first to divide the system into simpler parts, analyse them with the basic dynamical system representation of equation (2.1), then reconstruct the parts into a whole in order to make predictions. It is however increasingly appreciated that the *divide and conquer* approach fails short of making precise and yet significant or relevant statements about the system's behaviour as its complexity increases. A detailed characterisation of the underlying biochemical or biophysical mechanisms alone does not guarantee a deeper understanding of the reconstructed system. The dilemma is that although we recognise the limits of a reductionist approach we concede that there is no simple or intuitive alternative available. We recognise the fragmentary approach to reality in our mathematical model of a genomic analysis, using equivalence and fuzzy relations as the basis for comparisons and reasoning, but then aim to transcend the object-level by considering mappings between sets rather than the objects themselves. This transition is, in mathematics, at the heart of category theory.

2.6 EXAMPLE IV : GENETIC SYSTEMS

Genomics is the field of biological research taking 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, ultimatively, the organism. The two central questions are:

- ▷ "What do genes do?"
- ▷ "How do genes interact?"

As defined previously, system theory is a family of methodologies for the analysis of organisation and behaviour through mathematical modelling. A typical system theoretic approach to the two questions is to

- ▷ *Cluster* genes with known biological function according to similarity in pattern of *gene expression*[10].
- ▷ *Classify* genes with unknown function according to their similarity to the prototypes obtained from the clustering.
- ▷ *Identify* the parameters of a gene-network (dynamic) model using the cluster prototypes obtained previously.

The challenges for a system theoretic approach are:

- ▷ Very large number of variables (thousands of genes).
- ▷ Very small number of measurements (say between 8 and 18)
  - – repeated experiments usually not available.
  - – data often unreliable, missing, noisy or imprecise.
- ▷ Data are collected from a dynamic process under "closed-loop control".
- ▷ The processes usually are non-linear and time-variant.
- ▷ Information fusion of transcriptome and proteome data is non-trivial.

The first two items lead to the so called *dimensionality problem*. To this date, the majority of bioinformatics techniques have been concerned with the assembly, storage, and retrieval of biological information, with data analyses concentrated on sequence comparison and structure prediction. The move to functional genomics demands that both sequence and experimental data are analysed in ways that permit the generation of novel perspectives on gene and/or protein action and interaction. An approach to this problem is the construction of proper formal mathematical, parametric models that are

[10]Gene expression is the process by which a gene's coded information is converted into the structures present and operating in the cell. Expressed genes include those that are *transcribed* into mRNA and then *translated* into protein and those that are transcribed into RNA but not translated into protein (e.g., transfer and ribosomal RNAs).

identified from the data. As the focus in genomics is shifting from molecular characterisation to understanding functional activity, system theory is going to play an increasingly important role in providing biologists with better tools to extract information from data, as well as supporting new ways of thinking to characterise molecular systems in a general way, and quite independently of their physical and chemical constitution. The previous example on molecular modelling suggested that for more complex systems with a large number of objects (say thousands of genes), we require an approach that can integrate knowledge about the objects without physical or chemical interactions between individual genes being described in detail. What follows is an example for the approach we are going to develop in detail throughout the remaining part of this text.

Microarray technology provides us with gene expression measurements on the transciptome level. A typical experiment can provide measurements of the expression level of thousands of genes over a number of experimental conditions or over time. Considering a time-series of n samples, we can represent the observation of an individual signal (gene $u \in U$) as a point in the n-dimensional observation-space $X(f)$. Points that form a cluster have similar expression profiles and are subsequently postulated to have related biological function.

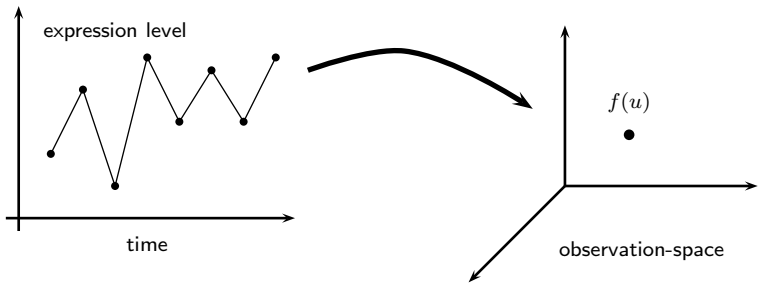


Fig. 2.7 From time-series to observation or factor-space representation.

Here the factor f denotes measurements on the transcriptome level. For a more complete picture of gene expression additional factors, for instance describing measurements on the proteome level, are introduced. As shown in Section 3.1, the factor-space approach extends naturally to several factors. A phenomena investigated refers to a specific biological concept C which we aim to characterise with the factors defined in Definition 2.3 (cf. Proposition 2.1). The extension of concept C in U is then the fuzzy mapping $\widetilde{A}$:

$$\begin{aligned} \widetilde{A}\colon \quad U &\rightarrow [0,1] \\ u &\mapsto \widetilde{A}(u) \ , \end{aligned}$$

where $\widetilde{A}(u)$ is the degree of relevance of u with respect to C or $\widetilde{A}$. When $\widetilde{A}(u) = 1$, u definitely accords with C, and for $\widetilde{A}(u) = 0$, u does not belong to $\widetilde{A}$ (a fuzzy attribute of C, i.e.,the function/expression of a gene in a specific context).

Clustering the points in the observation space $X(f)$, using partitional techniques such as the fuzzy-c-means algorithm [78], we are grouping genes (represented by measurements, i.e., points $f(u)$ in X) in order to infer the mapping $\widetilde{A}$ in U. Note that what we observe is a fuzzy set $\widetilde{B}$ on $X(f)$ (partition of X) and it is necessary to establish a relation between the 'model' $\widetilde{A}$ on U and the experimental evidence $\widetilde{B}$ in $X(f)$. The situation is similar to stochastic modelling and using descriptive statistics to approximate or estimate the model (parameters) from data. The fuzzy relational framework is intended to be a *theoretical* construct to complement *experimental* biology. The *biological principle* described is a conditional statement of the form

$$\text{IF } f(u) \text{ is } \widetilde{B}, \text{ THEN } C \text{ is } \widetilde{A} \ .$$

Let us have a closer look at the formal system described here. In Definition 2.3, a factor is defined as a mapping from a set of *abstract objects* $U \in U$ to space X. Here u denotes a gene, defined as a *conceptual entity* which exists apart from any specific encoding; it is that part of the natural system we wish to encode. Generalising the notion of a state in Example II, u is an ***abstract state*** of the natural system under consideration. Factor f evaluates the genes u in an experiment, leading to a numerical representation $x \in X(f)$. We note that any specific act of observation, experiment, is therefore at the same time an act of *abstraction*; theory and experiment are complementary and should not, cannot be separated.

In our scenario, illustrated in Figure 2.7, factor $f\colon U \to X(f)$ is a mapping from the set U of abstract states into an element of $X(f)$ which here is a point in the plane $\mathbb{R} \times \mathbb{R}$ of real numbers. Given any mapping between sets, the mapping f induces an ***equivalence relation*** E_f on its domain, by saying that $E_f(u_1, u_2)$ holds if and only if $f(u_2) = f(u_2)$. Therefore to say that the two genes u_1 and u_2 are related means that both produce the same 'effect' (observation) in our experiment.

If we form the quotient set U/E_f, we find that it is in one-one correspondence with the set of all possible values f can assume. This set, called *spectrum*, is denoted $f(U)$. If x is a point in $f(U) \subset X(f)$ we associate with x the entire equivalence class $f^{-1}(x)$. This means in effect that we can discuss the properties of our model (determined by an appropriate choice of factors f), in terms of the equivalence classes on U. As an important consequence, we have thus a means of comparing models or validating them with data. This important advance to the current practise of bioinformatics as we currently

lack conceptual frameworks that allow a formal analysis to which variables should be measured and why.

Applying clustering algorithms to the points in the observation space, we identify an (fuzzy) equivalence class $\widetilde{A}$ in U as a cluster of points in $X(f)$. Genes in U are grouped according to their similarity in expression profiles and hence allow us to predict their biological function. If we are to decide upon the similarity of two gene expression profiles by using the inequality $\|f(u_1)-f(u_2)\| \leq \varepsilon$ in the observation space, the inequality describes a subset (relation) $R_\varepsilon \subset U \times U$,

$$R_\varepsilon = \{(u_1, u_2) \in U \times U \ : \ \|f(u_1) - f(u_2)\| \leq \varepsilon\} \ .$$

This relation is not an equivalence relation, i.e., it is not a transitive relation. We can define a mapping $\widetilde{E}_\varepsilon$ such that $\widetilde{E}_\varepsilon(u_1, u_2)$ is greater than $1-\varepsilon$ if and only if u_1 and u_2 are indistinguishable with respect to the tolerance ε :

$$(u_1, u_2) \in R_\varepsilon \quad \text{if and only if} \quad \widetilde{E}_\varepsilon(u_1, u_2) \geq 1 - \varepsilon \ ,$$

where

$$\begin{aligned} \widetilde{E}_\varepsilon : \quad U \times U \ &\rightarrow \ [0,1] \\ (u_1, u_2) \ &\mapsto \ 1 - \inf\{\varepsilon \in [0,1] \ : \ (u_1, u_2) \in R_\varepsilon\} \end{aligned}$$

with $\varepsilon \in [0,1]$ and if there is no ε for which the relation holds, we define $\inf \emptyset \doteq 1$. $\widetilde{E}_\varepsilon$ is then a ***fuzzy equivalence relation***, also referred to as a *similarity relation*. The value $\widetilde{E}_\varepsilon(u_1, u_2) = 1 - \min\{|f(u_1) - f(u_2)|, 1\}$ describes the degree to which two objects u_1 and u_2 have similar observable consequences and transitivity of this relation implies that if u_1 and u_2 are similar and u_2 and u_3 are similar in their values in X, then u_1 is similar to u_3. Fuzzy equivalence relations will be further discussed in Section 3.5.

Fuzzy clustering algorithms return a matrix that specifies the degrees of membership of any u in the clusters (equivalence classes). We have seen, that the comparison of two real numbers with respect to an error bound ε induces fuzzy equivalence relations (a fuzzy set) and therefore suggests a fuzzy relational framework. There are however other reasons in support of a fuzzy mathematical approach. In many cases the evidence we have that a gene belongs to a cluster will be a matter of degree and w.r.t functional classes genes may belong to more than one class during an experiment.

By writing $f(u)$, the impression is that f is fixed and u is variable. However, the role of the argument and the mapping are formally interchangeable; we can keep u fixed and change the experimental setup. In which case, u becomes a mapping, whose arguments are themselves mappings: $\bar{u}(f) = f(u)$. The question "why $f(u)$?" can now be answered by "because u" or "because $\bar{u}$"

(cf. Proposition 4, Definition 4.2). Using fuzzy relations, the obtained formal system allows us to model causal entailment in natural systems (here gene regulatory networks).

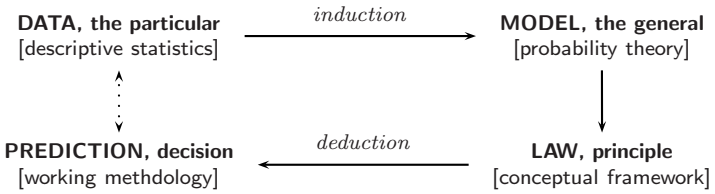


Fig. 2.8 The modelling process of a scientific investigation illustrating the difference of a conceptual framework and a working methodology. The square brackets refer to the example in the text.

To this point, we have discussed 'practical problems' but only 'in theory'. The aim for the rest of the text is to outline a conceptual framework for the study of gene-expression, gene-interactions and gene function. The relationship between such a *conceptual framework* and a *working methodology* can be explained by looking at the two complementary fields of statistics and probability theory. Using descriptive statistics, sample means, sample variances, histograms and relative frequencies, we extract information from data. On the other hand, a quantitative model based on random variables and probabilities, represents general relationships, going beyond the specific data set we may have, and is used to represent relationships which eventually describe natural laws or principles within a theory that captures the context of our scientific enquiry. In this respect statistics and probability theory, a sample mean and a mean, a unrelated. However, to justify a theory, model or principle, it should be possible to identify the model (its parameters) from experimental data. Only if both modelling pathways, the inductive step (system parameter identification) *and* the deductive step (model based predictions) are working to our satisfaction, the conceptual framework has explanatory value. A large part of statistics and probability theory is therefore devoted to the estimation and approximation of probabilistic concepts using statistics. Knowledge about the bias, variability and convergence of estimates makes us feel more confident in our conclusions. See Figure 2.8 for an illustration.

So why did we initially consider fundamental philosophical questions, when we are interested in genomics, a particular field of the biological sciences? It seems that many questions arising in philosophy have an analog in the sciences. The discussion of 'things as they are in themselves' (Kants world of phenomena) and the world of experience, of observable phenomena, is reflected in the modelling relation, i.e., in the process by which we model a natural system using formal mathematical objects. In the philosophy of science, the

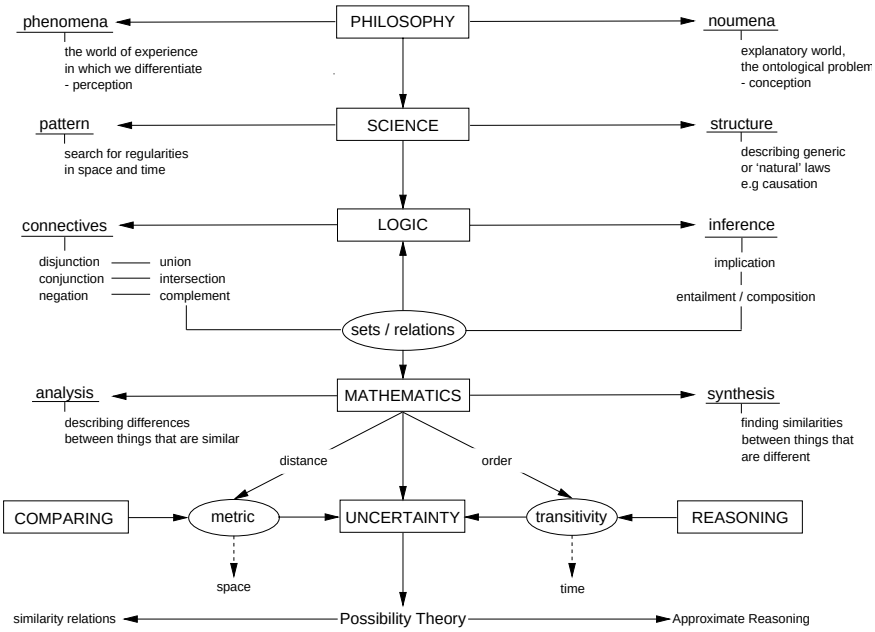


Fig. 2.9 "The fuzzy logic of scientific discovery". (From [77]).

problem of induction has been of particular importance. There seems now general consensus that the problem has no positive solution and that there is no single theory by whose means particular explanations could be conclusively shown to be true. In particular Karl Popper, tried to ensure that science, regardless of this apparent uncertainty, is put on a rational footing. Theories and hence models are worthwhile in that their comparison in applications, the verification with experimental data can generate new knowledge with an objective epistemic status. The philosophical problem of induction is in fact demonstrated by the problem of system identification, i.e., the estimation of model parameters from a finite set of data (the inductive aspect) and the use of the obtained model in forecasting (the deductive step). The philosophical position that scientific theories, extended beyond experimental data, cannot be verified in the sense of being logically entailed by them, suggests that we have have to pay particular attention to the representation of uncertainty in data, in models and in modelling. A philosophical investigation therefore gives us a bottom-up conceptual framework, providing reassurance, confidence and guidance in conduction scientific experiments and developing formal theories, models. Poppers view that unrefuted but corroborated hypotheses enjoy some special epistemic advantage, independent of anybody's attitude towards them, is confirmed by the common experience that we learn most from those models that failed.

Using the concepts discussed in the second example, in molecular biology, we may be able to model the interactions and relationships between say five genes with accuracy but we find it impossible to infer from this submodel the behaviour and function of the larger system in which it is embedded. In linguistics, we may be able to identify individual words of a poem, their origin, use and interpretation, but we find it rather difficult to understand the meaning of a the whole poem from knowledge of its parts. In mathematics, we can follow and check individual steps of a proof, establishing validity and truth of its parts, but do not necessarily understand the proof as a whole. These examples illustrate ***the curse of reductionism***. To proclaim *holism* as an alternative seems natural but unfortunately there seem hardly any formal holistic approaches that would overcome the problems of reductionism. Meanwhile *integrative approaches*, combining techniques and integrating the context in which the reasoning takes place seems a reasonable pragmatic step forward. Based on the framework outlined in this section, in subsequent sections we shall not attempt to model a biological phenomena 'as it is' but rather 'as we observe it'. Instead of modelling the physical structure or flow of energy using for example differential equations or thermodynamics, we strive to capture the organisation and information of observable biological phenomena. Using the words of Klir [33], it is increasingly recognised that studying the ways in which things can be, or can become, organised is equally meaningful and may, under some circumstances, be even more significant than studying the things themselves. This is of course the aim of system science, which I expect to play an increasingly important role in the interdisciplinary research problems in the life sciences.

3

A Factor Space Approach to Genomics

3.1 TAKING A SNAPSHOT

The relevance, applicability and importance of fuzzy set theory and fuzzy logic is generally linked to successful applications in the domain of engineering, especially where subjective notions have to be modelled and matched with abstract data structures. Examples of this include applications in the area of nonlinear control, expert systems and pattern recognition. The purpose of this section is to outline the conceptual foundations of a framework, based on the mathematics of fuzzy sets, that can be successfully employed to model some of the most complex phenomena in molecular biology.

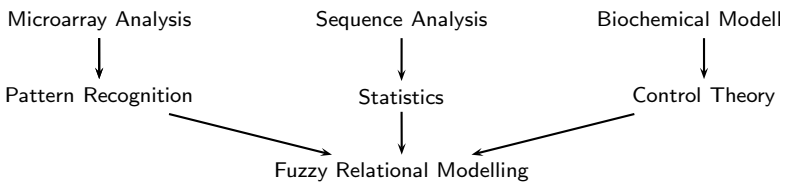


Fig. 3.1 The relationship of the fuzzy relational factor-space approach to biochemistry, bioinformatics and mathematical biology.

Our formal model describes a genome as a collection of genes. This set is equipped with a mathematical structure for logical inference. To allow rea-

soning in the presence of uncertainty, we need to formalise biological *concepts* and *facts* associated with these. Relationships between concepts and factors are expressed in terms of *rules*. Though the proposed mathematical language is 'in principle' complete – as accurate as biological knowledge is, a working methodology realistically is confined to specific aspects of a natural system. This *uncertainty principle* between the generality and predictive power of a model was summarised by Lotfi Zadeh[1] :

> "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."

3.1.1 Conceptual Framework vs Working Methodology

The dualism of *probability theory* and *statistics* is an useful analogy to illustrate the difference between a conceptual framework and a working methodology. The motivation for fuzzy relational biology is to create a conceptual framework in which problems of genome analysis can be formulated in the way many problems in science and engineering are translated into probability theory, i.e. formulated by means of random variables. Once this 'translation' has taken place, and is accepted as a reasonable *model* for the experimental context, we can reason about data and make predictions about events that have not yet been validated experimentally.

A key idea in probability theory is that of a random variable; which is neither random nor variable but simply a *mapping* from the sample space of elementary outcomes (providing evidence) to the event space in which we form our hypotheses. A random variable describes an *observable factor* of the experiment and is as such the 'real-world interface', relating experimental *outcomes* with theoretical *events*. However, probability theory itself does not consider experimental data. If we are to analyse measured or sampled data, we require statistics as a means to validate probabilistic concepts. Take for example the *concept* of 'central tendency' or 'mean value' which is defined abstractly using the expectation operator. In an experiment we use sample statistics to estimate or approximate these concepts in order to validate the formal model using random variables.

Although engineers and scientists frequently ignore or overlook the *modelling relation* between probability theory as a conceptual framework and the process which generates their measurements, consistent reasoning with exper-

[1]Lotfi Zadeh is widely regarded as the 'father' of fuzzy set theory. Although philosophers and mathematicians have long debated the inadequacy of 'crisp' mathematics in solving many real-world problems, it was Zadeh who initiated an avalanche of research into fuzzy mathematics, fuzzy logic, and possibility theory.

imental data using statistics as a working methodology, requires a conceptual framework that complements it. One advantage to have a conceptual framework in conjunction with a working methodology is the possibility to analyse properties of our model. It puts us in a position *to quantify model accuracy* and the *uncertainty of predictions*. In other words, a conceptual framework allows us *be precise about uncertainty* in our investigation.

3.1.2 Knowledge Representation: Conceptualisation

Let the *description frame* of a genome be denoted by $(U, \mathcal{C}, F)$, where $C \in \mathcal{C}$ denotes a concept and $f \in F$ describes a characterisation in terms of observable objects $u \in U$. We hereafter have two alternative cases 1) for the study of a single gene, it is represented as a concept $C \in \mathcal{C}$ while factors $f \in F$ describe different aspects of the expression or function of the gene; 2) studying large numbers of gene, for instance using microarray data, the context is denoted by C while the genes are the objects $u \in U$. For instance in yeast, respiration or fermentation could be the context in which all genes are studied. The result could then be a grouping (clustering) of the genes into these functional classes. We may therefore consider a gene as both an object or concept. Whichever situation is chosen it does not matter for the formal model. Since we wish to stress the fact that a gene is not a physical structure but a concept, we may call it a concept even if it is represented as an object in our formal model. We should also keep in mind that factors themselves are general in the sense that a factor should cover for a wide range of cases. For example, a factor may represent distances, positions, lengths, a gene's annotation (e.g its membership in a functional class), expression levels (e.g light intensity) or peptide masses. The form of the factors however has an effect on the formal model as we shall discuss further below.

The three ingredients $(U, \mathcal{C}, F)$ compose our formal model which is then built from data in the following way. An *object* u is either measured or verbally characterised with respect to a certain *factor* f. For example, u may be an ORF and $f(u)$, the *state* (e.g expression level) is a value in $X(f)$. $X(f)$ is referred to as the state space of factor f. The *relevance* of a *symptom* for a particular *phenomenon* is captured by a fuzzy set $\widetilde{B}$ in $X(f)$. The relevance of object $u \in U$ to the context or concept $C \in \mathcal{C}$ is expressed by the fuzzy set $\widetilde{A}$ in U. In general, we do not know $\widetilde{A}\colon U \to [0,1]$ *a priori*. The purpose of a model is to establish knowledge about $\widetilde{A}$, which describes a particular phenomenon, by means of observations $f(u)$ establishing symptom $\widetilde{B}$. The two-way relationship between the formalisation of a genome and the modelling of a particular aspect of it is summarised in Figure 3.2.

Genes are *functional entities* which cannot easily be defined physically. That is, genes are not simply a structural entity or DNA subsequence of

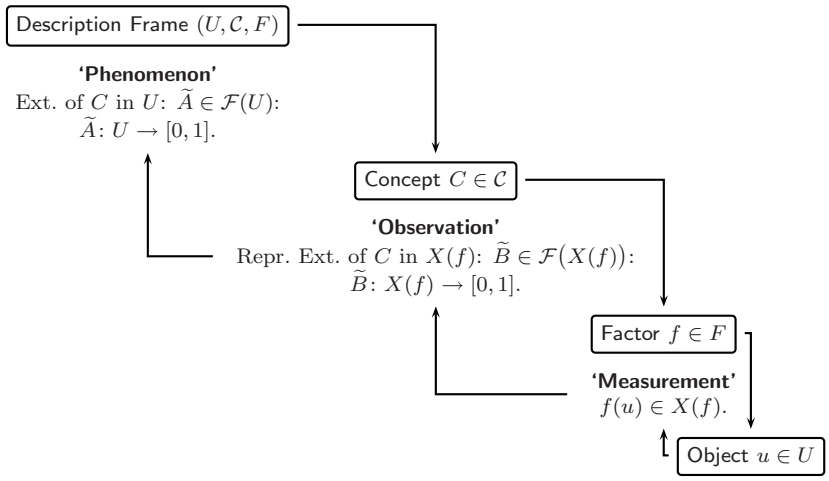


Fig. 3.2 The formal representation of a genome in terms of genes, factors and objects. The path following the framed boxes describes the key elements of the proposed conceptual framework, whereas the associated 'backward' path describes the working methodology representing gene expression and gene function from data.

the genome. We therefore view a gene as a *concept* characterised by various *factors*. Such a representation should be more integrated than the direct membership approach (which associates with elements in the genome a degree of relevance to the gene). For example, modelling gene expression, an example for two factors are measurements on the transcriptome level (mRNA synthesis) using for example microarrays and secondly measurements on the proteome level (cf. table 1.1). As for now, mass spectroscopy is much more accurate than gel analysis but is also far more laborious. It will nevertheless be important to study gene-expression on both levels as not all RNA is translated into protein. The factor-space model can provide a formal mathematical model of the interactions or relationships between those measurements in the presence of uncertainty.

Current biological genome analysis, by producing *maps*[2], establishes experimentally an "element–membership" description of the genome. This information is then in turn used to derive knowledge which establishes facts and their relationships. Instead of producing a sequence model and fitting experimental data to locations in the genome, we here propose a formalisation of the genome by its genes and their biological function or role. Although in

[2]In genome analysis, a *map* is a chart showing positions of genetic and/or physical markers in a genome.

this framework sequence information is not central, it is nevertheless vital. In the mathematical model developed in subsequent sections, sequence information can be used in studying the structure of genomes, for example, in the prediction of gene locations based on synteny[3].

In our model, knowledge is represented in terms of *concepts*, *objects*, and *rules*. In their mathematical representation the main tools are factor spaces and fuzzy maps. Factors are used to capture the characteristics of a concept, and fuzzy sets are used to describe the relationship (relevance, association, membership,...) of *objects* to a concept. The description of the essence and attributes of a concept by factors is known as *intension* while the aggregate of objects characterising a concept is known as *extension*. The extension of concept C is then an ordinary or fuzzy subset $\widetilde{A}$ of the objects universe U. What follows is an outline of the mathematical equipment we will use to take a picture...

The formalisation of a concept (conceptualisation of a gene) is based two aspects: *intension* and *extension* [73]. An extension of a concept C is an ordinary or fuzzy subset $\widetilde{A}$ of the objects universe U. These atomic data objects may for example be (sub)sequences, ORFs, and so forth[4]. Intension is defined by the collection of factors and their attributes characterising the concept. The classical definition of a set requires any genome subsequence to be either associated with the gene or not. In other words, given any open reading frame (ORF), for classical set theory we assign truth values 0 or 1 to define a *crisp set*. This application of the *law of the excluded middle* defines crisp sets on which we then build a bivalent logic. In many real world problems it is rather difficult to exactly decide whether an element has the property in question or not. In these cases where either the problem under consideration is a matter of degree or in which these decisions are subject to uncertainty, we use fuzzy sets and possibility distributions[5].

3.1.3 The Modelling Relation: Formalisation

As we study processes in a wide range of conditions, we find that there are *relationships* that remain effectively constant. These necessary relationships between objects, events, and conditions at a given time and those at later

[3] *Synteny* refers to a pair of genomes in which at least some of the genes are located at similar map positions.
[4] We use here the term *object* instead of *component*. An object can be a member of a set of independent abstract objects. Abstract objects may for example be answers to a question or concepts. Components are understood as interconnected objects or concepts. The factor $f: U \to X(f)$ as a mapping from U to $X(f)$ can also be understood as a component.
[5] A possibility distribution is in its formal definition identical to a fuzzy set. The semantics however differ. See [77] for more details.

time are what we call causal entailment or causal laws. We however note that the assumption of causality is usually accompanied by some form of abstraction, i.e., it may imply a simplification by conceptually taking the process considered out of its context, to ignore details and thereby achieving generality. As we almost never are able to include all influential factors in our model, principles must always be complemented by specifying the conditions and context in which we have found that they are applicable. The incompleteness of our model, the context induces uncertainty which we have to consider when drawing conclusions. Our concepts concerning causal relationships will then be true only relative to a certain approximation and to certain conditions. In this section we introduce the tools employed to describe the modelling relation between natural systems and a formal model. In table 3.1 biological concepts, their formalisation and considerations of the modelling relation are summarised.

Table 3.1 Summary of the formalisation of the 'modelling relation' for genome analysis.

	Problem	Formalisation
1.	Phenomenon: Gene function, gene expression	concepts
2.	Characterisation of 1) by means of observable facts	factors
3.	Structural components, or concepts	objects
4.	The *general* relationship between 1) and 2)	representation extension
5.	For a known, *particular* expression, relationship between 1) and 2)	feedback extension
6.	Representation of 1) by means of independent factors	G-envelope
7.	Precision of feedback extension and 6)	measure of coincidence

Definition 1 (Fuzzy Sets). Let all objects of a concept under discussion form a *universe* U. A *fuzzy set* $\widetilde{A}$ on the given universe U is defined by a mapping which associates with any object $u \in U$ a real number $\mu_{\widetilde{A}}(u) \in [0,1]$ in the unit interval, where $\mu_{\widetilde{A}}(u)$ is called the degree of membership of u in $\widetilde{A}$:

$$\begin{aligned} \widetilde{A}: \quad U &\rightarrow [0,1] \\ u &\mapsto \mu_{\widetilde{A}}(u) \ . \end{aligned}$$

The set of all fuzzy sets defined on U is denoted by $\mathcal{F}(U)$. For the sake of simplicity, we make no distinction between fuzzy set $\widetilde{A}$ and its membership function $\mu_{\widetilde{A}}$ and write $\widetilde{A}(u) \doteq \mu_{\widetilde{A}}(u)$. The definitions of this subsection are identical to those in [37] and [73].

Definition 2 (Factors). A factor f, is a common description of its *states* and its *characteristics*. An *object* u is relevant to a factor f if there exists a state $f(u)$ of f corresponding to U. Let U be a set of objects and V be a set of factors. The pair (U, V) is assumed to satisfy the condition that for any $u \in U$, V contains all factors relevant to u. Hence (U, V) defines a (crisp) *relation* R between U and V, where $R(u, f) = 1$ if u is relevant to f. We define

$$D(f) = \{u \in U \ : \ R(u, f) = 1\} \tag{3.1}$$
$$V(u) = \{f \in V \ : \ R(u, f) = 1\} \ . \tag{3.2}$$

A factor $f \in V$ is defined as the mapping

$$\begin{aligned} f\colon \quad D(f) &\rightarrow X(f) \\ u &\mapsto f(u) \end{aligned}$$

where $X(f) = \{f(u)\}$, is called the *state space* of f and $u \in U$.

Remark. Definition 2 may be generalised to allow for uncertainty in the knowledge about the relevance of an object u to a factor f. R is then defined as a fuzzy relation such that $\widetilde{R}(u, f) \in [0, 1]$.

Remark. We make a distinction between various types of factors. A factor may be *measurable* (the genes position in basepairs, width, ...) or *ordinal* (e.g degrees expressed in the unit interval $[0, 1]$). For *nominal* (categorical, qualitative) factors we can evaluate the equality for any two values $f(u) = f(u')$ as being either true or false. For example, a gene may be considered "functional" or "nonfunctional"[6]. Another example is the functional class (annotation) of a gene represented by a factor. On the other hand, for *cardinal* (non-nominal, quantitative) factors such as ratios or real-valued measurements, the comparison of two values may not be straightforward. The type of factor considered has implications on the mathematics as will be discussed further in Sections 3.5.1 and 3.5.4.

Without loss of generality, we extend the domain of f to the whole set U with the understanding that U is chosen to coincide with $D(f)$. We consider factors of a gene to be *observable* properties, either direct measurements of properties of sequence data or derived knowledge. A state is a sign or symbol that represents a special instance of a factor. When a state or a characteristic of a factor is used as the condition of producing certain results or effects, we

[6] A nonfunctional copy of a gene is also called a *pseudogene*.

say that the results of effects are attributable to the factor, not the state or the characteristic.

In general, an object is either a concept such as a gene or a structural element such as a segment of the genome, measures or characterises a sequence or is the measurement of some event in an organism. The latter corresponds then to the definition of an *observable* [61] in Robert Rosen's *modelling relation*[7] between a *natural system* and a *formal system* (illustrated in Figure 2.1). The affinity to Rosen's description of the modelling relation suggests that a factor-space approach to genome analysis may not only provide a "hands on" approach to data analysis but also a "heads on" conceptual framework to explain biological phenomena.

By accepting the existence of the modelling relation, factors become the means by which we encode and observe properties of the natural system under consideration. Using factors and representing them as mappings between the two spaces U and $X(f)$, we take the measurement and modelling process itself into account. As we shall see further below, this will allow us to be precise about model uncertainty, something other models avoid by hiding undesirable properties in assumptions about the natural system.

A factor f is equal to a factor g, if they are equal mappings, that is, $D(f) = D(g)$, $X(f) = X(g)$, and $f(u) = g(u)$ for any $u \in D(f)$. It is possible for states of a factor g to be a subset of the states of another factor f. A factor g is called a *proper subfactor* of f, denoted $f > g$, if there exists a (non-empty) set Y such that $X(f) = X(g) \times Y$. A factor g is called a *subfactor* of f, denoted by $f \geq g$, if $f > g$ or $f = g$.

Definition 3 (Factor Spaces). The family of state spaces $\{X(f)\}_{f \in F}$ is called a *factor space* on U if F, the set of factors, is a Boolean algebra. Therefore, for any $f, g \in F$,

$$X(f \vee g) = X(f - g) \times X(f \wedge g) \times X(g - f) \ .$$

The concept of a state space, given here, is akin to the same concept in control theory, the 'parameter space' in pattern recognition or the 'phase space' in physics (where factors are called observables [61]). The main difference is

[7]Robert Rosen's modeling relation, originally conceived as a conceptual device for clarifying the relationship between natural systems and structures created for understanding such systems, is presented as an epistemological method that not only subsumes the scientific method but extends to all of human intellectual activity where the acquisition and exploration of knowledge is of concern. As Rosen himself emphasised, the scientific method is a particular instance of a modeling relation.

that a factor space is more general than the usual assumption of an Euclidean or topological space. With the definition of a Boolean algebra, imposing a structure with intersection $\wedge$, disjunction $\vee$ and complement ^{c}, on F, we have a basis for logical reasoning with factors (and hopefully a tool for predicting biological properties and function).

Definition 4 (Conjunction, Disjunction of Factors). A factor h is called the *conjunction* of factors f and g, denoted by

$$h = f \wedge g$$

if h is the greatest common subfactor of f and g. In other words, $h = f \wedge g$, if and only if (iff) $X(h)$ is a common subspace of $X(f)$ and $X(g)$. Similar, a factor h is called the *disjunction* of factors f and g, denoted by

$$h = f \vee g$$

iff $X(h)$ contains subspaces of $X(f)$ and $X(g)$, and it is the smallest of such spaces. Both definitions apply to families of factors $g = \bigwedge_{i\in I} f_i$ and $g = \bigvee_{i\in I} f_i$ respectively.

Definition 5 (Independent-, Difference-, and Atomic Factors). Any two factors are called *independent* if their conjunction results in a *zero factor*, denoted $\mathbf{0}$, whose only state is the empty state. A factor h is called the *difference factor* between factors f and g, denoted by

$$h = f - g \ , \quad \text{if} \quad (f \wedge g) \vee h = f \quad \text{and} \quad h \wedge g = \mathbf{0} \ .$$

A factor f is called an *atomic factor* if f does not have proper subfactors except the zero factor. The factors in the set of all atomic factor are independent.

A zero factor is equivalent to the empty set in set theory. If a family of factors $\{f_j\}_{j=1,..,r}$ is independent, then

$$X\left(\bigvee_{j=1,..,r} f_j\right) = \prod_{j=1}^{r} X(f_j) \ . \tag{3.3}$$

Let V be a family of factors and let F be a set of factors of V such that F is sufficient, i.e., satisfying

$$\forall u, u' \in U, \ \exists f \in F \ : \ f(u) \neq f(u') \ . \tag{3.4}$$

To this point, we therefore assume that we have a sufficient number of factors describing a gene such that for a given object, there exists at least one factor

f in F, such that their state values differ in f. In Section 3.5 we will discuss the (possibly more realistic) situation in which a factor $f\colon U \to X(f)$ is a surjection[8]. That is, in an experimental context we may find that there are objects for which $f(u) = f(u')$, i.e., some objects are indistinguishable for f.

The triple $(U, \mathcal{C}, F)$ or equivalently $\left(U, \mathcal{C}, \{X(f)\}_{f \in F}\right)$ is called a *description frame* of $\mathcal{C}$ and is our formal representation of an experiment or investigation. Let $(U, \mathcal{C}, F)$ be a description frame and $C \in \mathcal{C}$. The *extension* of C in U is a fuzzy set $\widetilde{A} \in \mathcal{F}(U)$ on U, where $\widetilde{A}$ is a mapping :

$$\begin{aligned} \widetilde{A}\colon \quad U &\to [0,1] \\ u &\mapsto \widetilde{A}(u) \end{aligned} \tag{3.5}$$

where $\widetilde{A}(u)$ is the degree of relevance of u with respect to C or $\widetilde{A}$. When $\widetilde{A}(u) = 1$, u definitely accords with C, and for $\widetilde{A}(u) = 0$, u does not belong to $\widetilde{A}$ (a fuzzy attribute of C, i.e., the function/expression of a gene or a metabolic pathway). The fuzzy restriction $\widetilde{A}$ is therefore used to describe the phenomenon under consideration. Obviously, the crisp case, in which knowledge of the association of an object u with concept C is certain, $\widetilde{A}(u) = \{0, 1\}$, is a degenerate case of the given definition. The fuzzy mapping $\widetilde{A}$ defined on U is the ultimate aim of model as it describes the relationship of subsequences (ORFs') to a gene expression pathway. In general, we do not know $\widetilde{A}$ *a priori* but must establish knowledge about $\widetilde{A}$ via observable factors f where $f(u) \in X(f)$. For a given description frame $(U, \mathcal{C}, F)$, every state space $X(f)$ is called a representation universe and hence a factor space is just a family of representation universes of $\mathcal{C}$.

Consider a gene involved in a specific gene expression pathway from initiation of transcription to synthesis of functioning protein. This expression pathway is the means by which the genome specifies the content of the proteome[9]. In no organism is this biochemical signature entirely constant. Changes in genome activity (either transient or permanent) lead to *cellular differentiation*, the adoption by the cell of a specialised physiological role. In this respect, genes may either be active or inactive in transient regulation. Genome activity in a cell is activated by *signal transmission* – for example hormons stimulating the cell. There are a number of cases in which the signalling molecule does not interact directly with the transcription factor, but instead influences genome activity in an indirect manner. An example is provided by the *catabolic repression* system of bacteria (see [8], pg. 269). In short, intra-

[8] A *surjection* is a mapping that is *onto*: more than one object map into the same $f(u)$.
[9] The *proteome* is the complete protein content of a cell.

cellular and extracellular glucose levels control whether or not operons[10] are switched on. The glucose levels may then be "high" or "low" depending on the involvement of specific genes or polypeptides. Here f is a factor involved in the signal transduction pathway and u are polypeptides. The aggregate of factors describes the concept C, "catabolite repression", and the "glucose level", described by the extension of C in U, is a fuzzy set $\widetilde{A}$ (e.g. "low").

In other cases, gene segments (such as V28 and V29-1 coding for a part of the β T-cell receptor protein) must be linked to other gene segments from elsewhere in the locus before being expressed. If we consider the objects u to be gene segments and factors f to be specific processes in a pathway, the fuzzy set $\widetilde{A}$ on U models a physiological effect while the conjunction of segments is modelled using Definition 4 for the conjunction of factors. In analogy to approaches using the covariation of the nucleotide content of positions in RNA to predict which positions interact with each other, one can use the covariation in the occurrence of proteins to create a model of which proteins depend for their function on each other. Such information could be used to reconstruct metabolic pathways or signalling pathways [26].

Remark. The majority of cellular functions are a result of a combination of genes. It becomes therefore important to study the interrelationship of genes. At present our attention is directed at the expression of pathways for individual genes in terms of factors. The reason is that groups of genes, with the expression of one gene linked to that of another, should be more easily studied once the factor space model for individual genes is established. This is due to the fact that individual genes and their expression/function are described by fuzzy sets. Groups of genes can then be dealt with conventional fuzzy mathematics. For example, reasoning about a compound of genes (such as an operon in bacteria), employs triangular norms for the conjunction of fuzzy sets. Rule-based reasoning with extensions $\widetilde{A}$ of C in U is done by means *approximate reasoning*, employing fuzzy logic or fuzzy relations. We shall address approximate reasoning in Section 3.2.2. If on the other hand a large number of genes are assayed in a microarray experiment, the context (e.g diauxic shift in yeast) describes the concept while the set of all genes (respectively the ORFs) define the objects $u \in U$.

Knowledge about $\widetilde{A}$ is gathered via measurements or observations; $f(u)$, taking values in the *representation universe* $X(f)$. $\widetilde{A}$ is the *phenomenon* induced by data objects and $f(\widetilde{A})$ are its observable *symptoms*. Formally, $f(\widetilde{A})$ is referred to as the extension of C in $X(f)$. The mapping $\widetilde{A}\colon U \to [0,1]$

[10] An *operon* is a group of genes involved in a single biochemical pathway and expressed in conjunction with one another.

is to capture the essence of a gene's function (or its expression pattern). The extension of C in $X(f)$, $f(\widetilde{A}) \in \mathcal{F}\big(X(f)\big)$, describing the expression of gene C, is defined (using Zadeh's extension principle[11]) as follows :

Definition 6 (Representation Extension of C in $X(f)$). For a given description frame $(U, \mathcal{C}, F)$, let $C \in \mathcal{C}$ whose extension is $\widetilde{A} \in \mathcal{F}(U)$. For any $f \in F$, the extension of f to deal with fuzzy arguments is defined by

$$f(\widetilde{A}): \quad X(f) \rightarrow [0,1] \qquad (3.6)$$
$$x \mapsto f(\widetilde{A})(x) = \bigvee_{f(u)=x} \widetilde{A}(u) \ .$$

Then $f(\widetilde{A})$ is a fuzzy subset of the representation universe $X(f)$, $f(\widetilde{A}) \in \mathcal{F}\big(X(f)\big)$, where $f(\widetilde{A})$ is called the representation extension of C in the representation universe $X(f)$.

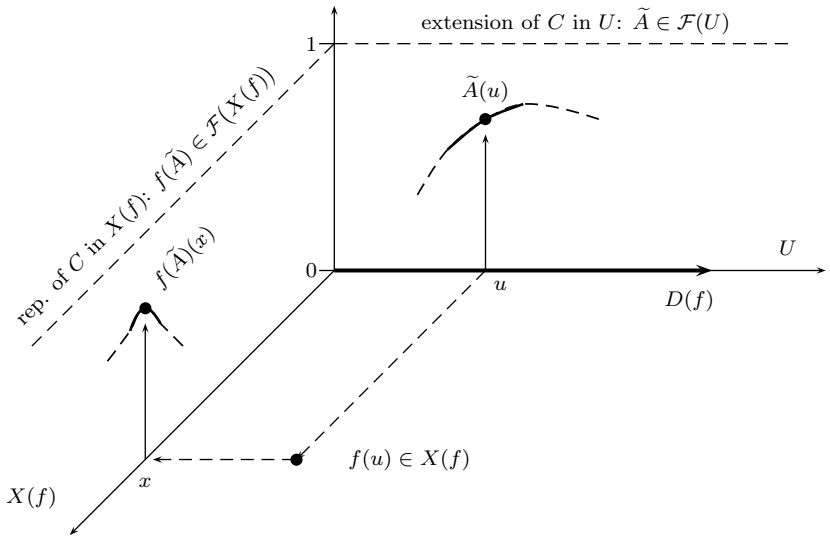


Fig. 3.3 Example of $(U, \mathcal{C}, F)$ for a specific $f \in F$ and $C \in \mathcal{C}$. The picture illustrates the relationship between the extension of C in U, $\widetilde{A} \in \mathcal{F}(U)$ and the representation extension of C in $X(f)$, $f(\widetilde{A}) \in \mathcal{F}\big(X(f)\big)$.

[11]The *extension principle* is a general principle by which a mathematical object, such as a function, can be *extended* to work for fuzzy sets.

The relationship between gene function (the phenomenon) and its characterisation by means of observable processes (gene expression) is therefore specified by $f^{-1}\big(f(\widetilde{A})\big)$, where $f(\widetilde{A}) \in \mathcal{F}\big(X(f)\big)$ and $f^{-1}\big(f(\widetilde{A})\big) \in \mathcal{F}(U)$. As a consequence of Zadeh's extension principle, used in the definition of the representation extension of C in $X(f)$ we have for any $u \in U$,

$$\begin{aligned} f^{-1}\big(f(\widetilde{A})\big)(u) &= f(\widetilde{A})\big(f(u)\big) \\ &= \bigvee_{f(u')=f(u)} \widetilde{A}(u') \geq \widetilde{A}(u) \end{aligned}$$

that is,

$$f^{-1}\big(f(\widetilde{A})\big) \supseteq \widetilde{A} \tag{3.7}$$

where equality is obtained for f being an injection (one-to-one mapping). Relation (3.7) therefore describes the quality of the model depending on the *model structure* – the choice of factors to model $\widetilde{A}$ on U. In [37] the following measure is introduced to quantify the *coincidence* of $f(\widetilde{A})(f(u))$ with $\widetilde{A}(u)$.

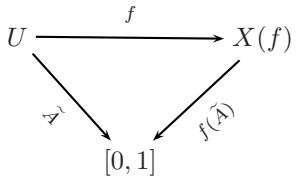
Definition 7 (Measure of Coincidence). Given a description frame $(U, \mathcal{C}, F)$, the mapping

$$\begin{aligned} \Lambda\colon \quad F \times \mathcal{F}(U) &\to [0,1] \\ (f, \widetilde{A}) &\mapsto \Lambda(f, \widetilde{A}) = \sup\Big\{1 - f(\widetilde{A})(f(u)) + \widetilde{A}(u)\colon u \in U\Big\}\ , \end{aligned}$$

is called the *measure of coincidence*. If we are to view a collection of factors as the *intension* of a concept C, the measure $1 - \Lambda(f, \widetilde{A})$ serves as a measures for the precision.

The relationship between the extension of C in U, $\widetilde{A} \in \mathcal{F}(U)$ and the representation extension of C in $X(f)$, $f(\widetilde{A}) \in \mathcal{F}\big(X(f)\big)$ is illustrated in Figure 3.3.

The essence of the *modelling relation* in Figure 2.1 is therefore captured by the mapping $f\colon U \to X(f)$. The *natural system* here is the concept of a gene, C, and is formalised by a fuzzy restriction $\widetilde{A}\colon U \to [0,1]$ on U. The basic assumption in modelling is that C can also be represented by means of observables in the factor space $\{X(f)\}_{f \in F}$:



We notice that for any factor $f \in F$, the inverse $f^{-1}\left(f(\widetilde{A})\right)$, which we shall discuss further below, is a composition of two mappings f and $f(\widetilde{A})$, that is,

$$f^{-1}\left(f(\widetilde{A})\right) = f(\widetilde{A}) \circ f \ .$$

and therefore

$$\widetilde{A} = f(\widetilde{A}) \circ f \ . \tag{3.8}$$

If $f^{-1}(x)$ is a single point set for every x in $X(f)$, then $f(\widetilde{A}) = \widetilde{A} \circ f^{-1}$ and for any state $x \in X(f)$, Definition 6 describes how we can define the fuzzy set $\widetilde{A}$ by the family of fuzzy sets $\{f(\widetilde{A})\}_{f \in F}$. With a family of independent factors, (3.3), our model may therefore also be seen as a (compound) rule ($f = \bigvee f_i$)

$$C\colon \text{ IF } f \text{ is } f(\widetilde{A}), \text{ THEN } u \text{ is } \widetilde{A} \ . \tag{3.9}$$

We shall pause for a moment in order to reflect how we have proceeded so far. We started of with a natural system described using observable factors which we represented by the mapping $f\colon U \to X(f)$. Any object $u \in U$ is consequently assigned a number, say in $\mathbb{R}$. As the objects are considered in a context, that is, with respect to a concept C, they induce a characteristic distribution (fuzzy restriction) $\widetilde{A}$ in U such that the relevance or association of u with C is quantified by $\widetilde{A}(u)$, a value in the unit interval. Since $\widetilde{A}$ is not known *a priori*, we gather *information* of concept C by means of observations or measurements in the image set (range) $X(f)$ of the factor f. Formally, we derive our knowledge in U via the representation extension of C in $X(f)$ leading to fuzzy restriction $f(\widetilde{A})$. In other words, our discussion of a concept C in terms of objects $u \in U$ has shifted to a discussion about the extension of a concept in U, $\widetilde{A}$ and its representation extension $f(\widetilde{A})$ in $X(f)$ or vice versa. Let us therefore look at the *fuzzy mapping* $\tilde{f}$, now from the set of fuzzy sets in $X(f)$ to the set of fuzzy sets in U :

$$\begin{aligned} \tilde{f}\colon \quad \mathcal{F}\big(X(f)\big) \ &\to \ \mathcal{F}(U) \\ f(\widetilde{A}) \ &\mapsto \ \tilde{f}\big(f(\widetilde{A})\big) = \tilde{f} \circ f(\widetilde{A}) \end{aligned} \tag{3.10}$$

where we can obtain $\mu_{\tilde{f}(f(\widetilde{A}))}(u)$ using the extension principle. For a family of independent factors $\{f_j\}$, let $X(f)$ be the Cartesian product of representation spaces $X(f) \doteq X(f_1) \times \cdots \times X(f_r)$, and $f_1(\widetilde{A}_1), \ldots, f_r(\widetilde{A}_r)$ be r fuzzy restrictions in $X(f_1) \times \cdots \times X(f_r)$, respectively. With f^{-1}, a mapping from $X(f)$ to U, $u = f^{-1}(x_1, \ldots, x_r)$, the extension principle defines a fuzzy restriction in U by

$$\widetilde{A} = \left\{ \big(u, \widetilde{A}(u)\big) \ : \ u = f^{-1}(x) \right\}$$

where

$$\widetilde{A}(u) = \sup_{(x_1,\ldots,x_r) \in f(u)} \min\left\{ f_1(\widetilde{A}_1)(x_1), \ldots, f_r(\widetilde{A}_r)(x_r) \right\} \ . \tag{3.11}$$

For $r = 1$, the extension principle reduces to

$$\widetilde{A}(u) = \sup_{x \in f(u)} f(\widetilde{A})(x) \ .$$

As with any mapping an equivalent representation for $\tilde{f}$ is the *fuzzy graph* defined by

$$\widetilde{\mathcal{G}} = f(\widetilde{A}_1) \times \widetilde{A}_1 \vee f_(\widetilde{A}_2) \times \widetilde{A}_2 \vee \cdots \tag{3.12}$$

or more compactly

$$\widetilde{\mathcal{G}} = \bigvee_{k=1} f(\widetilde{A}_k) \times \widetilde{A}_k \ ,$$

where the $f(\widetilde{A}_k)$ and $\widetilde{A}_k$, $k = 1, 2, \ldots$, are fuzzy subsets of $X(f)$ and U, respectively; each Cartesian product $f(\widetilde{A}_k) \times \widetilde{A}_k$ is in fact a *fuzzy relation* in $X(f) \times U$; and $\vee$ is the operation of disjunction, which is usually taken to be the union. In terms of membership functions we may then write

$$\widetilde{\mathcal{G}}(x, u) = \bigvee_j \big(f_j(\widetilde{A}_j)(x) \wedge \widetilde{A}_j(u)\big)$$

where $x \in X(f_j)$, $u \in U$, $\vee$ and $\wedge$ are any triangular T- and T-conorm, respectively. The concept of a fuzzy graph is illustrated in Figure 3.4.

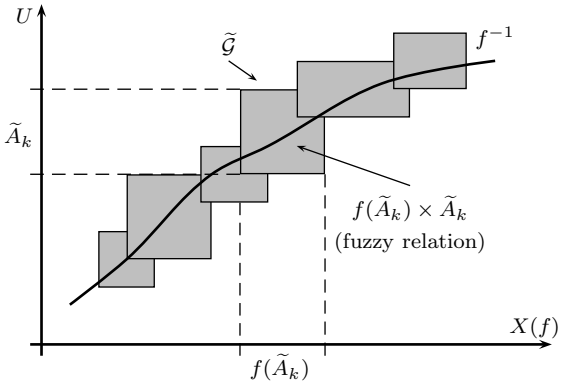


Fig. 3.4 Fuzzy graph $\widetilde{\mathcal{G}}$ as a composition of fuzzy relations.

3.1.4 The Modelling Relation: Reasoning About Data

Let $(U, \mathcal{C}, F)$ be a description frame an $\widetilde{A}$ be the extension of $C \in \mathcal{C}$. In biological terms $(U, \mathcal{C}, F)$ represents the experiment, $\widetilde{A}$ the function of a gene, the phenomenon investigated with respect to $C \in \mathcal{C}$ while $f(\widetilde{A}) \in \mathcal{F}\big(X(f)\big)$

models the expression of C as observed. According to the definition of factor spaces, for any factor $f \in F$, the representation extension $f(\widetilde{A})$ of C in the representation universe $X(f)$ is determined by $\widetilde{A}$. As argued before, $\widetilde{A}$ will usually be unknown but instead we may know the representation extension $\widetilde{B}(f)$ of C in $X(f)$. For example, the success of ORF scanning depends on the frequency with which termination triplets appear in the DNA sequence. Knowledge about the frequency of occurrence and codon bias combined with the typical length of ORFs will be expressed in $\widetilde{B}(f)$. What follows is a discussion of how the extension of $\widetilde{A}$ in U can be determined from $\widetilde{B}$. In other words, in analogy to (3.7), we need to establish the relationship between the *general* – the phenomenon under consideration and a given or known *particular* symptom (observation). The concept of *feedback extension* developed by Li et al. [37] is the appropriate tool to study this relationship.

Definition 8 (Feedback Extension of C w.r.t f). Let $(U, \mathcal{C}, F)$ be a description frame with $C \in \mathcal{C}$ and $f \in F$. Assume $\widetilde{B}(f)$ to be a known representation extension of the concept C in the representation universe $X(f)$. The *feedback extension* of C with respect to f is defined by

$$\begin{aligned} f^{-1}\big(\widetilde{B}(f)\big): \quad U &\to [0,1] \\ u &\mapsto f^{-1}\big(\widetilde{B}(f)\big)(u)\ . \end{aligned} \tag{3.13}$$

Then $f^{-1}\big(\widetilde{B}(f)\big)$ is a fuzzy subset of the universe U.

For any $f \in F$, if the representation extension of a concept C in $X(f)$ is known, then

$$\begin{aligned} f\left(f^{-1}\big(\widetilde{B}(f)\big)\right)(x) &= \bigvee_{f(u)=x} f^{-1}\big(\widetilde{B}(f)\big)(u) \\ &= \bigvee_{f(u)=x} \widetilde{B}(f)(f(u))\ , \end{aligned}$$

and hence

$$\underbrace{f\Big(\overbrace{f^{-1}\big(\widetilde{B}(f)\big)}^{\text{feedback extension}}\Big)}_{\text{representation extension}} \subseteq \widetilde{B}(f) \tag{3.14}$$

which becomes an equality when f is a surjection (the mapping is onto: more than one u can map into the same $x \in X(f)$).

Expressions (3.7) and (3.14) mean that our models of $\widetilde{A}$ in U will generally be an approximation "from above". Note also that the injection condition under which equality is obtained in (3.7) may be too restrictive to describe

gene expression. However, for a model based on synteny, f describing the location of an ORF, the mapping will indeed be injective and our description of $\widetilde{A}$ by means of a family of fuzzy sets $f(\widetilde{A})$, $f \in F$ is, in principle, accurate.

Definition 8 provides the basis for modelling gene function by means of observable gene expression. The following issues require further discussion in the biological context. This discussion can be based on the theoretical results mostly found in [37]:

1. For all $f, g \in F$, if $f \geq g$, we have
$$f^{-1}\left(f(\widetilde{A})\right) \subset g^{-1}\left(g(\widetilde{A})\right) \ .$$
meaning that a more complicated factor describes the expression of a gene more accurately than with a simpler (dependent) subfactor.
2. Decomposition of a factor into a set of simpler (independent) factors.
3. The representation extension of family of simpler factors and the accuracy in describing the extension of a concept in U.

Definition 9 (Cylindrical Extension). Let $\{X(f)\}_{f\in F}$ be a factor space on U. Assume that $f, g \in F$ with $f \geq g$, and any $\widetilde{B} \in \mathcal{F}\big(X(g)\big)$. The *cylindrical extension* of $\widetilde{B}$ from g to f is a fuzzy subset of $X(f)$:

$$\begin{aligned} \uparrow_g^f \widetilde{B}\colon \quad X(f) &\ \rightarrow\ [0,1] \\ (x,y) &\ \mapsto\ \left(\uparrow_g^f \widetilde{B}\right)(x,y) = \widetilde{B}(x)\ , \end{aligned} \tag{3.15}$$

where $X(f) = X(g) \times X(f-g)$, $x \in X(g)$, $y \in X(f-g)$ and $\uparrow_g^f \widetilde{B} \in \mathcal{F}\big(X(f)\big)$.

Using cylindrical extension we can get an approximate representation extension $\uparrow_g^f g(\widetilde{A})$ of C, w.r.t a more complicated factor f, from the representation extension $g(\widetilde{A})$ of C w.r.t to a simpler subfactor g :

$$f^{-1}\left(\uparrow_g^f g(\widetilde{A})\right) \supset f^{-1}\left(f(\widetilde{A})\right)\ . \tag{3.16}$$

Since the approximation of the feedback extension (3.16) of a more complex factor by means of the feedback extension of a simpler factor leads to inaccuracies, we shall now discuss the modelling of an extension $\widetilde{A}$ on U by means of independent factors.

Theorem 1 (Li [37] pg. 64). *Let $(U, \mathcal{C}, F)$ be a description frame and $\widetilde{A}$ be the extension of $C \in \mathcal{C}$. For any $f, g \in F$, for $f \wedge g = \mathbf{0}$,*

$$f^{-1}\left(f(\widetilde{A})\right) \cap g^{-1}\left(g(\widetilde{A})\right) = (f \vee g)^{-1}\left(\left(\uparrow_f^{f\vee g} f(\widetilde{A}\right) \cap \left(\uparrow_g^{f\vee g} g(\widetilde{A})\right)\right) \ .$$

From theorem 1, let $G \subset F$ where elements of G are mutually independent then

$$\widetilde{A}[G] \doteq \bigcap_{f\in G} f^{-1}\left(f(\widetilde{A})\right) \tag{3.17}$$

is called the *G-envelope* or *G-feedback extension* of $\widetilde{A}$. The G-envelope (3.17) approximates the extension of $\widetilde{A}$ by independent factors such that for example with two factors $f, g \in G$, we have $\widetilde{A} \subset f^{-1}\left(f(\widetilde{A})\right) \cap g^{-1}\left(g(\widetilde{A})\right)$, i.e., the phenomenon is approximated from above. For $f = \bigvee_{i\in I} f_i$, we have $X(f) = \prod_j X(f_j)$. Since $\widetilde{A} \subset \widetilde{A}[G]$ and $u \in U$, we can write

$$\widetilde{A}[G](u) = \bigwedge_j f_j\left(\widetilde{A}\right)\left(f_j(u)\right) \ . \tag{3.18}$$

Using the coincidence measure $\Lambda(f, \widetilde{A})$ from Definition 7 on page 59, with $\Lambda' = 1 - \Lambda(f, \widetilde{A})$ we can define a measure of how precise our approximation is. Li et al. [37] provided an inequality for Λ' determined from individual factors :

$$\Lambda' = 1 - \Lambda(f, \widetilde{A}) \leq \bigwedge_{j=1}^{r} \Lambda'_j$$

where

$$\Lambda(f, \widetilde{A}) = \Lambda\left(\bigvee_{j=1}^{r} f_j, \widetilde{A}\right) \geq \bigvee_{j=1}^{r} \Lambda\left(f_j, \widetilde{A}\right) \ .$$

The smaller Λ', the higher the degree of coincidence of the intension with its concept C.

In practical applications, the extension $\widetilde{A}$ will usually be unknown. If we put $\widetilde{B}(f_j) = f_j(\widetilde{A})$, the precision with which the gene function is described by means of observed gene expression is

$$\begin{aligned} \widetilde{A}(u) \leq \widetilde{A}[G](u) &= \bigwedge_{j=1}^{r} \widetilde{B}(f_j)(x_j) \\ &= \left(\prod_{j=1}^{r} \widetilde{B}(f_j)\right)(x_1, \ldots, x_r) \end{aligned} \tag{3.19}$$

where $x_j \doteq f_j(u)$. Expression (3.19) describes how we form the representation extension $\widetilde{B}(f_j)$ of the concept C on the representation universe $X(f_j)$. The observed symptoms $\widetilde{B}(f_j)$ are identified from measured sample data and/or context dependent expert knowledge. Let $f = \bigvee_j f_j$, given the representation extension $\widetilde{B}(f_j)$, we construct cylindrical extensions $\uparrow^f_{f_j} \widetilde{B}(f_j)$. The intersection of cylindrical extension is then our approximate representation extension of C on the representation universe $X(f)$:

$$\bigcap_j \left(\uparrow^f_{f_j} \widetilde{B}(f_j) \right) \supset \uparrow^f_{f_j} \widetilde{B}(f) \ . \tag{3.20}$$

Figure 3.5 is to illustrate the approximation, which is in fact a subset (not shown) of the "pyramid" in the product space $X(f_1) \times X(f_2)$.

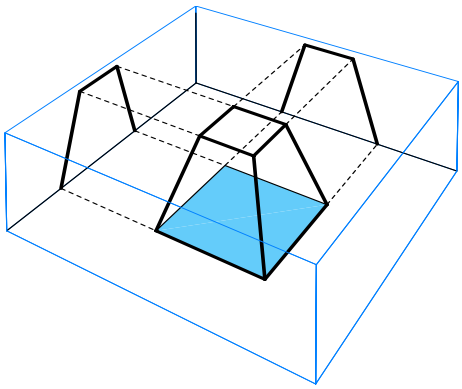


Fig. 3.5 Approximation of the representation extension of concept C in $X(f)$ by two independent factors (describing the two axis). The approximation is obtained from the representation extensions $\widetilde{B}(f_j)$ of C in f_j via their cylindrical extension into the product space and intersection.

Definition 10 (Direct Product of Fuzzy Sets). Let U and U' be two universes and $\widetilde{A} \in \mathcal{F}(U)$ and $\widetilde{A}' \in \mathcal{F}(U')$, respectively. We can form a new fuzzy set, called *direct product* (Cartesian product) of $\widetilde{A}$ and $\widetilde{A}'$, denoted $\widetilde{A} \times \widetilde{A}'$, whose membership function is defined by

$$\left(\widetilde{A} \times \widetilde{A}' \right)(u, u') = \widetilde{A}(u) \wedge \widetilde{A}(u') \qquad \forall (u, u') \in U \times U' \ .$$

From (3.19), (3.20) and with Definition 10 for the direct product of fuzzy sets, we find that $\widetilde{A}[G]$ can be described by the intersection of representation

extensions $\widetilde{B}(f_j)$:

$$\begin{aligned}\widetilde{A}[G] &= \left(\bigcap_{i=1}^{r} \widetilde{B}(f_j)\right) \circ f \\ &= \left(\bigcap_{j=1}^{r} f_i(\widetilde{A})\right) \circ f \ .\end{aligned} \tag{3.21}$$

By choosing an appropriate intersection operator we can achieve different approximations. A well known class of operators, called T-norms, is discussed further below. With the representation extension $\widetilde{B} = f(\widetilde{A})$ defined in terms of the representation extension $\widetilde{B}(f_j)$,

$$\widetilde{B} \approx \prod_{j=1}^{r} \widetilde{B}(f_j) \ , \tag{3.22}$$

we have for any $u \in U$ the approximation

$$\begin{aligned}\widetilde{A}(u) &\approx f(\widetilde{A})(f(u)) \\ &\approx \bigwedge_{j=1}^{r} \widetilde{B}(f_j)(f_j(u)) \ .\end{aligned} \tag{3.23}$$

The approximation (3.23) of the modelling relation may also be represented as a if-then rule :

$$\text{IF } f_1(u) \text{ is } \widetilde{B}(f_1) \text{ AND } f_2 \text{ is } \widetilde{B}_2 \text{ AND } \cdots \text{ AND } f_r \text{ is } \widetilde{B}_r, \text{ THEN } C \text{ is } \widetilde{A} \ . \tag{3.24}$$

Remark. Note that the assumption of independent factors implies a loss of information. Given a set of genes $u \in U$, we want to answer questions of the form "Is u expressed?". In general, we ask whether u manifests, realises or instantiates a property $\widetilde{A}$? That is, we want to treat $\widetilde{A}$ as a predicate or adjective, of a give referent u. We want to determine the truth, validity of some (synthetic) proposition $\widetilde{A}(u)$ about u. Equation (3.23), is an example of reductionism in our model. It represents a kind of fractionation of an arbitrary predicate or property representing $\widetilde{A}(u)$ as a conjunction of essentially independent subproperties $f(\widetilde{A}_j)$ ($\widetilde{B}(f_j)$ respectively). The claim made is that we can recapture the property $\widetilde{A}$ from the expressions observed on subspaces by purely syntactic means. The purpose of the assumption is of course that is allows a reasonably simple implementation of the approach in a soft- or hardware.

As described by equation (3.8), given the description frame $(U, \mathcal{C}, F)$, $C \in \mathcal{C}$, gene function $\widetilde{A} \in \mathcal{F}(U)$ is modelled by a set of characteristic variables

(factors) for which $\{f(\widetilde{A})\}_{f\in F}$ describes the expression of gene C such that

$$\widetilde{A} = f(\widetilde{A}) \circ f \ . \tag{3.8}$$

Then, given the observed expression $\widetilde{B}(f) \in \mathcal{F}\big(X(f)\big)$, from the feedback extension of C with respect to f, we evaluate the statement "C is $\widetilde{A}$", as

$$\widetilde{A}' = \widetilde{B}(f) \circ f \ . \tag{3.25}$$

The composition is read as "first apply f and then $\widetilde{B}(f)$",

$$\widetilde{B}(f)\big(f(u)\big) = \big(\widetilde{B}(f) \circ f\big)(u) \ .$$

Equation (3.25) is akin to the *compositional rule of inference*, forming the basis for *approximate reasoning* which we shall look at in the following section.

Remark. In this section, the role of fuzzy restrictions (fuzzy sets) has been the representation of semantic *information* (or equivalently uncertainty), either *observed* or induced from *measured* data. It is important to realise that with factor-space theory, we represent a biological process or an organism in terms of its *components*, their *function* and their interaction by *relations* via the *information* they provide. In this respect our model differs considerably to conventional physical models of organisms. In the factor-space model, a factor represents a component or part of the system fulfilling its function as defined by the mapping that associates objects with some observable consequence. An important extension to this still reductionist perspective is the information captured by fuzzy restriction $\widetilde{A}(u)$ – with respect to some *context* (concept) of the 'whole' (the genome or description frame). The 'whole' is thus present in the part by 'constraining' not the part itself but the information it carries. In other words the extension from mappings (factors) to associated fuzzy sets allows us to capture external influence on constituent parts of the whole (system).

3.2 IMAGE ANALYSIS

In the field of fuzzy mathematics (fuzzy logic or fuzzy systems), equation (3.25) is also known as the *compositional rule of inference* for *approximate reasoning* [77, 78]. This equivalence is important as it should provide us with ways to identify the fuzzy graph (3.12) from data. This section should therefore be considered together with the results from page 60.

3.2.1 Black and White Negatives: Classical Two-Valued Logic

The basis of *propositional calculus* is a set of formal entities, that is, simple statements - primitive propositions, often called *variables* of the logic. Such

variables are combined using basic logical connectives, $\vee$ (or), $\wedge$ (and), $\neg$ (not), to build expressions or *production rules*. The mapping from the set of all expressions into the set of truth-values is called *truth-evaluation*. In two-valued logic the 'truth' of a proposition can take the values 0 ("false") and 1 ("true") only. The formula $B \Rightarrow A$, modelling "IF B THEN A" or "B implies A" is called *material implication* and is defined by

$$B \Rightarrow A \doteq \neg B \vee A \tag{3.26}$$
$$= (B \vee A) \vee \neg B \ , \tag{3.27}$$

that is, it is defined in terms of the three basic connectives. From the table above it becomes apparent that two-valued logic is insufficient to deal with all those cases for which we might employ rule-based knowledge. In particular, the truth values of B and $B \Rightarrow A$ cannot be chosen independently and it is not possible to quantify gradual changes in the antecedent and consequent of a rule. In fact, if the antecedent B were interpreted as the cause and the consequent A as the effect, the material implication would mean that an absent cause entails any effect. Further, every proposition implies itself as $B \Rightarrow B$, meaning everything is self-caused.

A propositional calculus is a logic of atomic propositions which cannot be broken down. The validity of arguments does not depend on the meaning of these atomic propositions, but rather on the form of the argument. If we consider propositions of the form "all a's are b" which involves the *quantifier* "all" and the *predicate* b, then the validity of an argument should depend on the relationship between parts of the statement as well as the form of the statement. In order to reason with this type of proposition, propositional calculus is extended to *predicate calculus*. A predicate on a set is a relation. Generalising the Boolean modens ponens, we can either redefine the implication or allow fuzzy concepts (fuzzy sets) in the premise and conclusion parts of the rule. This leads to what is known as *approximate reasoning*.

3.2.2 Approximate Reasoning: Compositional Rule of Inference

In approximate reasoning, classical propositions B, A, which can either be true or false are replaced by fuzzy propositions such as "f is $\widetilde{B}$" where f is a fuzzy variable and the fuzzy concept $\widetilde{B}$ is represented by a fuzzy set. A given fact "f is $\widetilde{B}'$", conjunctively combined with the prior knowledge of the implication rule, leads to gradual truth values taking values on the unit interval. In standard logic the emphasis is on formal validity and truth is to be preserved under any and every interpretation. On the other hand, in approximate reasoning one tries to preserve information within the situation (context) in which the reasoning takes place. In general we identify the triple $(\neg, \wedge, \vee)$

with $({}^c, \cap, \cup)$ and here in particular with $(1-\mu, T, S)$. Here, a T-norm[12] is a binary function that extends the domain of logical conjunction from the set $\{0,1\}$ to the interval $[0,1]$. Similarly, S models disjunctive operations.

A *proposition* takes the form "f is $\widetilde{B}$" with fuzzy variable f taking values in X and $\widetilde{B}$ modelled by a fuzzy set defined on the *universe of discourse* X by *membership function* $\mu\colon X \to [0,1]$. A *compound statement*, "f is $\widetilde{B}_1$ AND g is $\widetilde{B}_2$", is taken as a fuzzy set $\widetilde{B}_1 \cap \widetilde{B}_2$ in $X_1 \times X_2$ with

$$\mu_{\widetilde{B}_1 \cap \widetilde{B}_2}(x_1, x_2) = T\big(\mu_{\widetilde{B}_1}(x), \mu_{\widetilde{B}_2}(x)\big)$$

For the sake of simplicity we consider a single rule of type

$$\text{IF } f \text{ is } f(\widetilde{A}), \text{ THEN } C \text{ is } \widetilde{A}$$

which can be regarded as a fuzzy relation

$$\begin{array}{rccl} \widetilde{R}: & X \times U & \to & [0,1] \\ & (x,u) & \mapsto & \widetilde{R}(x,u) \end{array}$$

where $\widetilde{R}(x,u)$ is interpreted as the strength of relation between x and u. Viewed as a fuzzy set, with $\mu_{\widetilde{R}}(x,u) \doteq \widetilde{R}(x,u)$ denoting the degree of membership in the (fuzzy) subset $\widetilde{R}$, $\mu_{\widetilde{R}}(x,u)$ is computed by means of a *fuzzy implication*.

Replacing the negation $\neg$ in (3.26) with the basic fuzzy complement $1-\mu$, and the disjunction $\vee$ with the fuzzy union max-operator, we obtain the so-called *Dienes-Rescher implication*

$$\widetilde{R}(x,u) = \max\big(1 - f(\widetilde{A})(x), \widetilde{A}(u)\big)\ . \tag{3.28}$$

From (3.27), replacing negation by the fuzzy complement, disjunction by the max-operator and conjunction by the min-operator, we obtain the *Zadeh implication*

$$\widetilde{R}(x,u) = \max\Big(\min\big(f(\widetilde{A})(x), \widetilde{A}(u)\big), 1 - f(\widetilde{A})(x)\Big)\ . \tag{3.29}$$

Other possibilities are :

$$\widetilde{R}(x,u) = \min\big(1, 1 - \widetilde{A}(x) + f(\widetilde{A})(u)\big) \qquad : \text{Lukasiewicz implication,} \tag{3.30}$$

$$\widetilde{R}(x,u) = \begin{cases} 1 & \text{if } f(\widetilde{A})(x) \le \widetilde{A}(u), \\ \widetilde{A}(u) & \text{otherwise.} \end{cases} \qquad : \text{Gödel implication.} \tag{3.31}$$

[12] Triangular or T-norms are further discussed in Section 3.5.1.

Or defined using T-norms

$$\widetilde{R}(x,u) = \min\big(\widetilde{A}(x), f(\widetilde{A})(u)\big) \quad : \text{Minimum implication,} \tag{3.32}$$

$$\widetilde{R}(x,u) = f(\widetilde{A})(x) \cdot \widetilde{A}(u) \quad : \text{Product implication.} \tag{3.33}$$

Finally, given some 'input data', the generalised *modus ponens* provides a mechanism for inference on the basis of $\widetilde{B}$:

Implication:	IF f is $f(\widetilde{A})$, THEN C is $\widetilde{A}$.
Premise:	f is $\widetilde{B}$.
Conclusion:	C is $\widetilde{A}'$.

In terms of fuzzy relations the output fuzzy set $\widetilde{A}'$ is obtained as the relational sup-t composition, $\widetilde{A}' = \widetilde{B} \circ \widetilde{R}$. The computation of the conclusion $\widetilde{A}'(u)$ is realised on the basis of what is called the *compositional rule of inference*. The inference can be described in three steps as illustrated in Figure 3.6 :

1. Extension of $\widetilde{B}$ to $X \times U$, i.e., $\widetilde{B}_{\text{ext}}(x,u) \doteq \widetilde{B}(x)$.
2. Intersection of $\widetilde{B}'_{\text{ext}}$ with $\widetilde{R}$, i.e., $\widetilde{B}_{\text{ext}} \cap \widetilde{R}(x,u) = T\big(\widetilde{A}_{\text{ext}}(x,u), \widetilde{R}(x,u)\big) \forall (x,u)$.
3. Projection of $\widetilde{B}_{\text{ext}} \cap \widetilde{R}$ on U, i.e., the compositional rule of inference is defined by

$$\begin{aligned} \widetilde{A}'(u) &= \sup_{x \in X} \widetilde{B}_{\text{ext}} \cap \widetilde{R}(x,u) \\ &= \sup_{x \in X} T\big(\widetilde{B}_{\text{ext}}(x,u),\ \widetilde{R}(x,u)\big) \end{aligned} \tag{3.34}$$

where in (3.34) the suprenum (or maximum for a finite representation universe) can be seen as a 'selection' from the information provided by $\widetilde{B}_{\text{ext}} \cap \widetilde{R}$. Taking the maximum over all values in X, one may view $\widetilde{A}'$ described by $\widetilde{A}'(u)$ as the *shadow* of fuzzy set $\widetilde{A}'_{\text{ext}} \cap \widetilde{R}$.

Remark. As the name suggests, the compositional rule of inference is related to the composition of (fuzzy) relations. The symbol $\circ$ denotes the composition of two functions defined as follows. Given any two functions, g and h, when the codomains of g is the domain of h, as in

$$X \xrightarrow{\ g\ } \Omega \xrightarrow{\ h\ } U$$

The *composite function* $h \circ g$ is defined as the set of ordered pairs

$$\big\{(x,u) \ :\ x \in X, u \in U, \text{ and } \exists\ \omega \in \Omega \text{ with } (x,\omega) \in g \text{ and } (\omega,u) \in h\big\} \tag{3.35}$$

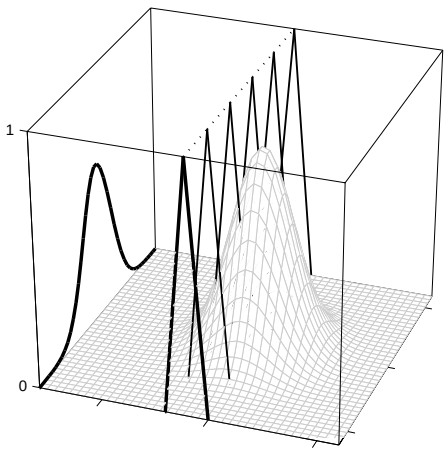
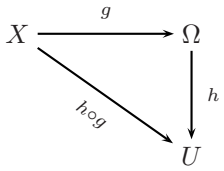


Fig. 3.6 Compositional rule of inference in approximate reasoning.

Illustrating the composition with the commutative diagram



we can interpret the composition as the rule "first apply g, then apply h", $(h \circ g)(x) = h(g(x))$, formalising the idea of two operations carried out in succession. Now let g and h define two ordinary relations on $X \times \Omega$ and $\Omega \times U$ respectively. The composition of g and h, denoted $h \circ g$, is defined as a relation in $X \times U$ such that $(x, u) \in h \circ g$ if and only if there exists at least one $\omega \in \Omega$ such that $(x, \omega) \in g$ and $(\omega, u) \in h$. Using characteristic function $\zeta_g \colon X \times \Omega \to \{0, 1\}$ and $\zeta_h \colon \Omega \times U \to \{0, 1\}$, we have

$$\zeta_{h \circ g}(x, u) = \max_{\omega \in \Omega} T\big(\zeta_g(x, \omega), \zeta_h(\omega, u)\big) \tag{3.36}$$

for any $(x, u) \in X \times U$ where T is any T-norm. Equation (3.36) is then generalised to fuzzy relations by simply replacing the characteristic function for crisp sets ζ by the fuzzy set membership function μ :

$$\mu_{h \circ g}(x, u) = \max_{\omega \in \Omega} T\big(\mu_g(x, \omega), \mu_h(\omega, u)\big) \tag{3.37}$$

Because the T-norm in (3.37) can take a variety of formulas, we obtain for each T-norm a particular composition.

After this excursion into approximate reasoning, we now return to our factor-space model by reminding ourselves that on page 60, we considered the fuzzy mapping (3.10) between representation space $X(f)$ and U as a fuzzy graph (3.12), where each *fuzzy point* in $X(f) \times U$ is in fact a fuzzy relation. The discussion on approximate reasoning is therefore directly applicable to the description of fuzzy graph (3.12) with only minor changes to notation[13].

The fact that factor-space models are firmly based in fuzzy mathematics and its application to rule-based systems, should provide us with a rich source of results and applications in building a working methodology. For example, the identification of fuzzy relations from numerical data is discussed in [57]. Another important aspect of the factor-space model may be the description of gene expression in terms of (fuzzy logic) if-then rules. The *interpretability* of the factor-space model may provide an 'interface' to context-dependent expert knowledge provided by biologists in addition to numerical data.

3.3 IMAGE ENHANCEMENTS

Amidst the avalanche of data describing genes and proteins it is evident that the dynamics of biological regulatory mechanisms cannot be understood by merely identifying components, cataloguing products and by drawing diagrams depicting how regulatory components are connected. A frequently discussed feature of natural systems is that of *emergent properties*. A system, composed of a multitude of simpler components may produce as a whole a higher degree of functional complexity. This seems to apply equally to a swarm of fish or birds as well as to the atoms that make up molecules, and molecules building an organism. Instead of studying the material structure of the components, it seemed to me therefore more interesting to investigate their interrelations and interactions. Mathematical relations are an obvious language to capture those concepts. While describing or recognising patterns (cluster, subsets,...) requires transitive equality relations and set operations such as union and intersections, to describe interactions (causal entailment) we require a more complex system model (cf. Fig. 2.9). Depending on whether we ask "What are the genes/proteins function?" or "How do genes/proteins interact?", we end up employing methods from pattern recognition or system modelling respectively. For example, using clustering algorithms to group data, in pattern recognition a classifier is build for classification. Describing interactions, possibly changing through time, a parametric model is build using methods from system identification. The purpose of the system model is then to a) explain unknown relationships or to make predictions.

[13]In the present section, we have simplified the notation $\widetilde{B}$ for $\widetilde{B}(f)$; $\widetilde{R}$ for $\widetilde{R}_f$; X for $X(f)$ and note that $\mu_{\widetilde{R}}(x,u) \equiv \widetilde{R}(x,u)$.

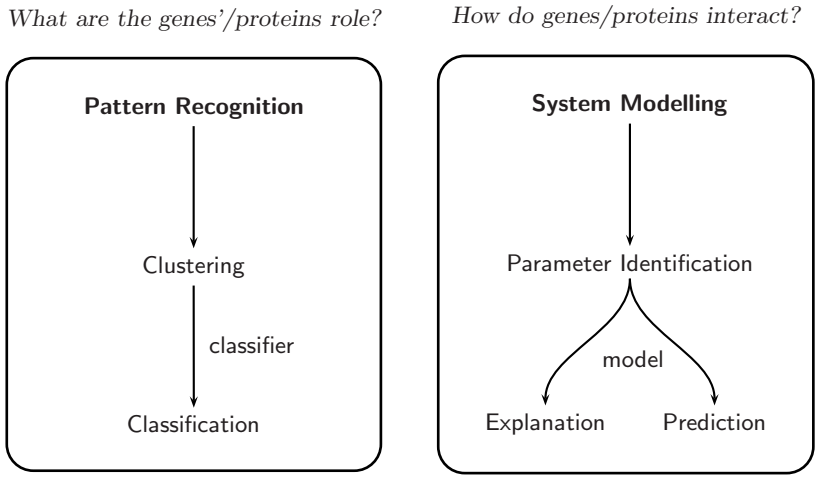


Fig. 3.7 Depending on whether we ask for a genes/proteins function (interrelationships between groups of objects) or whether we wish to describe interactions, mathematical (equivalence) relations form the basis for a formal model.

The aim of the present text has been to develop a formal mathematical framework to study gene expression, regulation and function. Apart from a conceptual framework, the objective is to outline a working methodology, which could be applied to gene expression data. There are principally two reasons why the fuzzy relational factor-space approach is promising. To this point, we should have fulfilled two principal requirements for a conceptual framework:

- It is 'rich' in structure such that it can deal with various situations arising from *complexity*.
- It is sufficiently 'complete' in that it is possible to deal with *uncertainty*.

A more detailed analysis of the capabilities of our approach is to follow in Section 3.6 after we have reviewed system theory used in biology and, in Section 3.5, have introduced more 'machinery' required for a qualified judgement.

To implement the proposed model in software we require access to data in form of information about experimental results (*observations*) stored in databases, and *measurements*[14] obtained directly from experiments, for instance, using gene microarrays. A particular genome information system that may be suitable is the GIMS project [53] developed at the University

[14]The distinction between 'measurements' and 'observations' will be discussed further below. See also Figure 3.12.

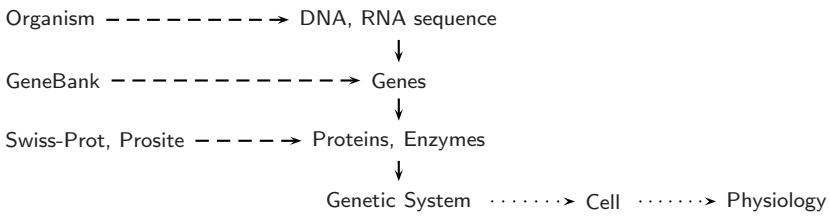


Fig. 3.8 The wider context in which the proposed project is considered.

of Manchester. GIMS (Genome Information Management System), focuses on genome and related data. Information is presented interactively by browsing and graphical displays and graph structures. The data are currently stored in a database accessed through Java. By combining information about coding sequences and gene-expression data we should be able to achieve much more accurate quantitative predictions but in general we will also obtain a more comprehensive understanding of gene function. As an example for the application in helping biologist in answering their question we refer to [39] where a combined algorithm is used for genome-wide prediction of protein function.

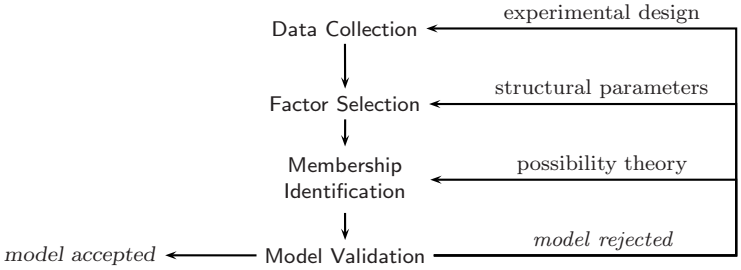


Fig. 3.9 Outline of model building and validation.

We can identify the following packages of further work required to expand and validate the concept. They are grouped according to the two main components of the proposal :

1. **Working Methodology:**
   (a) Analyse genomic data resources.
   (b) Validate the approach using sample data from public data libraries. For example, the GIMS database can be used to generate such data.
   (c) Construct representative test sets of sample data, in various degrees of completeness to determine the levels of genomic and functional information which can reliably be inferred using the model.

In addition to coding sequences, other non-coding segments such as regulatory elements and transposons should be examined and matched with models for a known genome.

(d) Validate the use of the compositional rule of inference in fuzzy relational systems for a realisation of the fuzzy mapping $\tilde{f}$.

(e) Provide a software implementation of the model.

2. **Conceptual Framework:**

(a) For the representation extension $\widetilde{B}(f)$ in $X(f)$, investigate techniques that estimate the membership function $\widetilde{B}(f)$ from sample (sequence) data (i.e., set-valued statistics, fuzzy clustering, pattern recognition,...)

i. Investigate mathematical properties of the feedback extension (Definition 8) and its accuracy if the representation extension $\widetilde{B}$ is identified from sample sequence data.

ii. Study the mapping represented by the rule IF f is $\widetilde{B}$, THEN C is $\widetilde{A}$. Discuss f in the context of Rosen's modelling relationship.

(b) On the basis of theorem 1 develop a measure of representation of C by a set of factors. Then for any two genes devise an algorithm that for a given extension $\widetilde{A}$ can determine the presence or absence of a particular gene in a pathway or genomic function.

(c) Describe a rule-based reasoning scheme based on the factor space representation of a genome.

(d) Study the decomposition of factors into simpler factors and consequences for the approximation of the extension $\widetilde{A}$ of C in U.

(e) Reformulate the description frame $(U, \mathcal{C}, F)$ in terms of category theory, such that it coincides with the framework outlined by Rosen [59, 60, 61].

(f) Could a factor f also be used to model Mendel's *genotypic factor* of heredity? In this case, the extension of C in $X(f)$ is the *phenotype*, and $\widetilde{A}$ on U the *genotype*.

(g) Investigate factor space models for the processing of molecular measurements and its possible link to genetic networks.

(h) Discuss the encoding or representation of time.

3.4 MOVING PICTURES

At present, the biological function of the unknown gene is inferred from good matches to genes of known function, based on the assumption that they share

a common biological ancestor from which they have evolved. Current practice is based on heuristics, mainly exploiting homology relationships, and there does not exist a formal mathematical framework to the identification of gene function from sequenced genome data. A formal model would aid gene function prediction by exploiting known higher-order genomic features such as synteny (gene order) and co-membership of known metabolic, regulatory or developmental pathways.

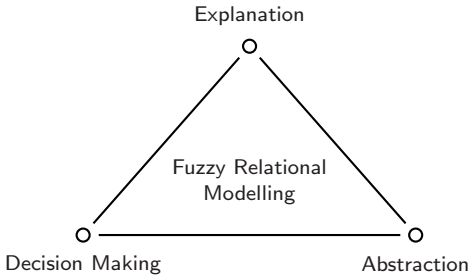


Fig. 3.10 Three purposes of modelling in genomic analysis.

Modelling organisms, their components, function and behaviour, can take different forms[15]. We may distinguish three main forms of modelling a) the attempt to understand the workings of a cell or organism, in order to gain some insight into "what actually happens", requiring *explanatory* models; b) for classification and prediction based on data, the model is motivated by *decision making*; c) given the knowledge of particular cases we wish to generalise; trying to find an *abstraction* from individual objects to classes of models or systems. Associated with these modes of modelling we have the areas of *simulation*[16], e.g. the mass-action rate models or energy models of molecules for which concentrations are our observables; *pattern recognition* as in sequence analysis where we identify patterns to for example predict protein structure. The tools employed are for instance neural networks, principal component analysis, clustering and so forth. None of these techniques seems to provide direct inside into the actual principles at work (causal entailment).

The area of 'bioinformatics', to this date, is largely engaged in the development of tools and technologies that support scientists in their explanation of biological phenomena. The internal structure of models is usually less relevant as long as the predictions or patterns produced by the model are suggestive.

[15] "What you see depends on how you look at it."
[16] What I referred to as modelling with the purpose of *simulation*, is equivalent to Rosen's concept of a *metaphor*: the decoding, prediction from a formal model without specific encodings (cf. Fig. 2.1). Examples for metaphors are Catastrophe Theory, Automata Theory, Chaos Theory and various other paradigms. The problem with these concepts is that by given up encoding, we also give up *verifiability* – ignoring in some way experimental data.

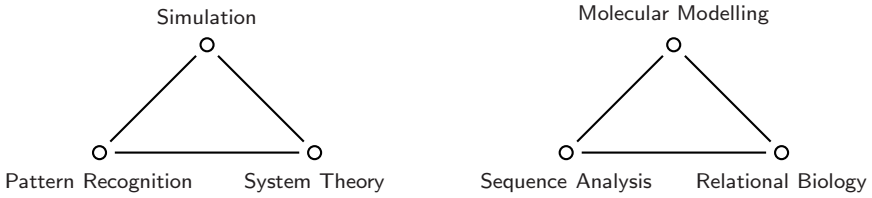


Fig. 3.11 Methodologies (left) and their applications (right) corresponding to the three 'modes' of modelling (explanation, decision making, abstraction).

A "this is why" answer is produced by the biologist and the abstract model he has in his mind not the formal mathematical model he may have used as a tool towards his aims. As more is learned about genomes, it will become increasingly important to have formal concepts which more directly generate insights, that currently are dependent on the reasoning skills of the scientist. At best, simulations provide us with an *"how"* but we have yet to produce formal models that explicitly demonstrate a *"why"* to the questions biologists ask.

How useful a model or theory is, how well it can explain unknown relationships, depends on the context in which it is used and what it is used for. Take for example evolutionary theory, probability theory or game theory. As the term 'theory' suggests, they do not apply to 'the particular' but are generalisations[17]. Probability theory does give you an impression what, on average, you could expect to occur when throwing an ideal dice. Considering the particular dice you hold in you hand, the theory will not be very be useful in predicting the outcome. Similar, evolutionary theory explains processes over a very long period of time, involving large groups and should not be used to explain individual human behaviour. The value of these theories lies in simulation and from it, the explanation of general patterns. If a simulation confirms expectations, appears 'realistic', one may be tempted to think that the components and structure that generated the simulation is a model of 'the thing itself'.

Idealisation, for instance viewing genes as switches, is useful and appropriate if later one can show that the pattern identified on the basis of these assumptions would hold if the idealisations are removed. Since "on/off", "present/abscent" idealisations are literally false, we should make sure not to succumb the temptation to extrapolate from the findings, e.g talking of "computations in cells". In other words, the implicit uncertainty of a model is im-

[17]In contrast to a *generalisation* (going from the particular or special to the general), *abstraction* does not imply a loss in the capability to describe particular aspects of a system under consideration. An abstraction makes a problem context independent by transferring it into an alternative domain.

portant and should be stated explicitly. Therefore if we cannot build models that represent gene expression levels and enzyme catalytic activities as graded changes, we may produce simulations using some idealisations but should keep in mind the consequences. Idealisations are useful to build tractable models of large or complex systems; for instance using automata theory to model autocatalytic networks as done by Stuart Kauffman [29]. These models, like game theory applied to economics, simulate processes and help us to extract patterns (general principles) that govern the process as a whole. Here the idealisation works as long as we do not make predictions for a particular component, at a particular instant of time or in a particular condition. The foregoing discussion akin to the philosophical discussion of Section 2. Although we cannot know the noumena itself, we can know about it in the phenomena. Instead of striving for a detailed model of the process itself, we should try to capture the process by which biologists, using informal models, successfully generate new knowledge from measurements and observations.

3.4.1 Systems Theory in Molecular Biology

Although there have been various applications of systems theory to biology [40] and many of the concepts evolved into what is now known as *metabolic engineering*, attempts describe genetic systems fail at 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."

It has been suggested that the failure of these original applications to molecular biology is likely to be due to a naive transfer of Newtonian physics to biological processes. The view that organism or cells can be described in terms of energy or masses with forces acting on them, leads to accurate models only for small submodels loosing its predictive power for more complex systems. The key element of a Newtonian approach is that it views the cell as a material system to be analysed as a family of constituent parts. These objects define a state space for which dynamical relationships are specified. Such an abstraction yields a set of coupled differential equations (cf. Section 2.4).

Phenotypes are what we can observe directly about organisms. They are tangible, material properties that we can measure, can compare and experiment with. The phenotype is seen as being 'caused' or 'forced' by the genotype. As Rosen [62] points out, the phenotype–genotype dualism is allied to the Newtonian dualism between states and the forces that change the states. In Aristotelian language, the states represent material causation of behaviour, while the forces are an amalgam of formal and efficient causation. In recent years it has become possible to measure or observe and describe this relationship on the molecular level.

Remark. The word 'causality' is ambiguous. In Section 2 we introduced the "causal problem" in Schopenhauer's philosophical setting. With regard to the modelling relation (Section 3.1.3) causal entailment referred to causation in the phenomenal world on one hand and inferential entailment in formal systems on the other. The first comprehensive theory of causation was Aristotle's. It distinguishes four types of cause: the material cause (or stuff), the formal (formative) cause (or shape), the efficient cause (or force) and the final cause (or goal). For a formal logical system, given an 'effect', say proposition P, axioms correspond to the material cause of P, production rules are understood as the efficient cause of P and the specification of particular sequences of production rules or an algorithm is identified as the formal cause. For a dynamic system a state can itself be entailed only by a preceding state. If for a chronicle $\{(n, f(n))\}$ we ask *why* the n^{th} entry gives the particular value $f(n)$, the answer is *because* of the initial condition $f(0)$, i.e., $f(0)$ is the material cause; and *because* of a state transition mapping T for which $f(n+1) = T(f(n))$, i.e., T corresponds to the efficient cause; and *because* of exponent n from which $f(n)$ is obtained by iterating the transition map n times beginning with $f(0)$; i.e., n refers to the formal cause. In Rosen's relational biology, for a component $f\colon A \to B$, such that $a \mapsto f(a)$, the question "why $f(a)$?" is answered by "because f" and "because a". In other words, "a entails $f(a)$" or formally $f \Rightarrow (a \Rightarrow f(a))$. Here f corresponds to the efficient cause of ("effect $f(a)$"), and a refers to the material cause of $f(a)$. One of Rosen's achievements is that he introduced a formalism rich enough in entailment, to allow final causation without implying teleology.

Biological phenotypes, considered as material systems, are open. They are open to 'forcing' by genes as well as open to interactions with their environment. To study an open system it is therefore necessary to consider the "outside", the environment in order to understand what is going on "inside". A critic of such a Newtonian approach to biology and the failure of reductionism to supply the whole from its parts was Robert Rosen [59, 60, 61, 62]. In the Newtonian realm a system is 'closed' by internalising external influences through added state variables and more parameters to the system. Since Newtonian mechanics is a paradigm for mechanisms in general, it is worthwhile looking at the simplest dynamical system of a single particle moving under the action of a force [47] (see also page 35). The motion is governed by Newton's Second Law, which *defines* the force F acting on a mass point m to be the rate of change of momentum $(m \cdot v)$:

$$F = m \cdot \frac{\mathrm{d}v}{\mathrm{d}t} \;=\; m \cdot \frac{\mathrm{d}^2 x}{\mathrm{d}t^2}$$

where v denotes velocity which, in turn, is defined as rate of change of position or displacement from some origin of coordinates. Here 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:

$$F(x, v) = -\theta \cdot x$$

where θ is a parameter. We may refer to the response of the system to forces, as described by these equations, as the 'inertial' aspect while the exertion of forces by the system is corresponds to the system's 'gravitational' aspect.

Rosen's suggestion [62] is to shift attention from exclusively 'inertial', i.e., structural aspects such as the DNA molecule and its sequence, to 'gravitational' concepts. In other words, instead of concerns with material causation of behaviour, manifested in state sets, he suggested formal and efficient causations as the focus of attention. Such a shift of perspective is exemplified in category theory, Rosen's preferred language to discuss these problems in the abstract, by studying mappings between sets (of objects) rather than analysing the objects themselves. His compelling arguments for such a move and the formalism he provides open up a new dimension for the study of biological phenomena. Rosen considers the Newtonian conception responsible for a lack of progress in mathematical biology and argues the case for a new approach, called *relational biology*. He emphasises that we must look for principles that govern the way in which physical phenomena are organised, principles that govern the *organisation* of phenomena, rather than the phenomena themselves. Relational biology is therefore about organisation and describes entailment without states. The association of energy or matter, described by states and dynamical laws, is to be replaced by the description of a system in terms of its components, their function and contribution to the organisation of the system.

Such a transition of levels in formal analysis is exemplified by statistical mechanics. Considering a gas molecule, using Newtonian mechanics, the position of a molecule is specified by three spatial coordinates while the representation of the momentum or velocity require three further coordinates. Therefore, each molecule's position and momentum at any instant can be described by six coordinates. For N molecules a $6N$-dimensional phase-space captures the state of the system of N molecules as a point. Changes over time are described by a trajectory in this phase space. To perform exact calculations for a system of N molecules is virtually impossible and statistical mechanics provided an useful abstraction when it replaced the exact trajectory of the system state with probability density functions over regions of the phase-space. The discussion of any specific point in this space is 'replaced' by a mapping from the underlying reference set to the unit interval. Like in the factor-space model we have suggested here, we have surrendered absolute certainty and precision and given an approximate description where regions of the mathematical space which represents the natural system is characterised by mappings into the unit interval, expressing (un)certainty in modelling and data. The change of 'description mode' we suggest, away from a biochemical or structural repre-

sentation to representation of entailment, is akin to the change from classical physics to quantum mechanics. In microscopic physics, we are prevented by the nature of things from being able to ascertain the location and velocity of a particle at one and the same instance and therefore cannot predict with certainty the systems future state. Quantum mechanics subsequently abandons the application of causal connections. The question of whether or not the causal connection is true 'in reality', becomes somewhat irrelevant and to abandon the causal structure one also abandons the mechanistic view of systems in favour of a statistical concept which is mathematical.

Translating the description of any particular system, say the one we see in front of us in the laboratory, into a formal conceptual framework, we then consider a general class of systems to which our specific one is equivalent. Probability theory is not only the basis of statistical mechanics it is a conceptual framework in its own right (cf. Sec. 3.1.1, pg. 48) to which descriptive statistics provides a 'real-world interface'. In fuzzy relational biology we aim at an integration of synergy of a number of frameworks to provide a powerful approach to the challenges provided by genomics and proteomics. Once we have translated our problem at hand into the conceptual framework of our choice, we use its rules of inference to make predictions, simulations and ideally find explanations for the principles by which data and pattern are generated. Matching experimental reality with a theory, validating a mathematical model with data, will usually require us to average or aggregate data so as to reduce uncertainty. In the process our conclusions may suffer confidence and a molecular biologist may rightly be sceptical about the value of the efforts in creating a formal model. The only consolidation may be that we often learn most from those models which fail because the process building a formal model, as a way of thinking, is valuable in itself, providing a 'systematic' procedure supporting the hypothesis testing biologists do 'naturally'.

At the heart of relational biology is the modelling relation between a natural and formal system (Figure 2.1). The formal system describes a set of components, interrelated in a particular way. Any two natural systems that realise this formalism are considered as analogous, realise or manifest a common *organisation*. Consequently, the concept of a *realisation* and the building of realisations take a central role as they provide the mechanism that relates the formal model with the natural system in its physical appearance. The philosophical discussion in Section 2 (see also the summary in Section 3.6) provided the basis for objective knowledge in the realm of the phenomenal world. For the relational biology presented here are therefore two propositions of importance :

- ▷ Observable events can be represented by the evaluation of factors on abstract states (objects).

▷ A factor can be regarded as a mapping from a set of objects to a set of labels.

Note that in the fuzzy relational factor-space approach we generalise the concept of an observable in the Newtonian spirit to allow for qualitative characterisations of data objects (non-numerical observations). Rather than modelling the organisation of physical processes on a molecular level, our fuzzy relational model provides a phenomenological theory of gene function. We use the term *measurement* in a narrower sense, measuring or counting – quantifying characteristics of data objects, while we also allow derived knowledge to describe data objects. The latter is then referred to as an *observation*. Figure 3.12 outlines the multi-levelled representation of biological information and processing of data in the proposed fuzzy relational factor-space model and its possible extension to relational biology.

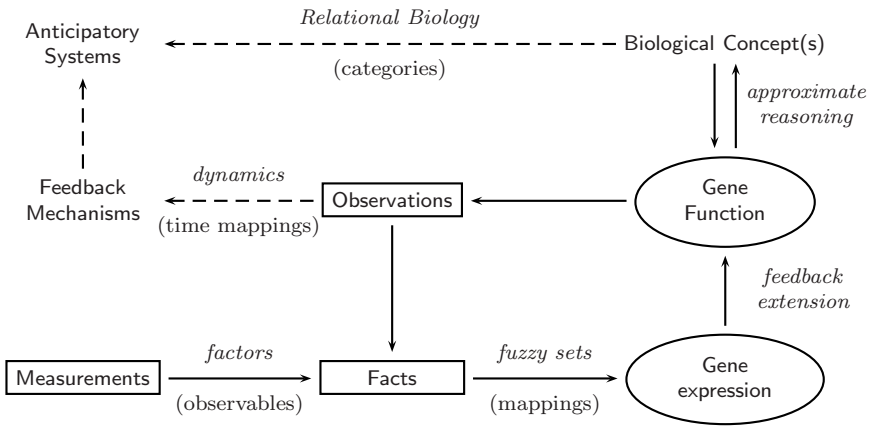


Fig. 3.12 Multi-levelled representation of biological information and processing of data in the proposed fuzzy relational factor-space model for genome analysis.

The fact that a fuzzy relational factor space genomic model is based on sets, mappings and functions, has an attractive consequence. Rephrasing our approach in terms of category theory would provide us with a *language of models*. This should eventually enable us to compare and study models not simply in terms of their 'predictive power' but also with respect to their principal capability to capture the essence of biological phenomena. However, as often we have to be careful not to overstate our expectations (and the required category theory isn't going to be easy either). Abstraction, as a shift of domain in which a problem is discussed, has been a key idea in the present text. Our first step was to translate an empirical or experimental problem into a description frame $(U, F, \mathcal{C})$ consisting of concepts $C \in \mathcal{C}$ (e.g. genes itself or a gene expression experiment), characterised by factors $f \in F$ which

evaluate objects $u \in U$. We have thus established a formal *system*[18] encoding a natural system. Another example of abstraction was in Section 3.1.3 the shift of the discussion of concepts in terms of objects to an analysis of the concept in terms of its extension (representation extension respectively). In Section 3.5, we will introduce another perspective of factors in terms of the equivalence relations they induce on U. Instead of a purely functional and abstract model we then have a means to validate or identify the model with experimental data.

Remark. The work of Rosen and the foregoing discussion suggest that basic questions of biology cannot only be solved empirically but can also be discussed in a conceptual framework. Recent development in the life sciences and genomics in particular, provide evidence for Rosen's contention that many biological problems are conceptual rather than empirical. The "post-genome" challenge is to be able to interpret and use the genome data: focus is shifting from molecular characterisation to understanding functional activity. (Fuzzy) relational biology is concerned with function and behaviour rather than structure. It is the system scientist's main occupation to understand ways and means of how to encode natural systems into formal systems by means of modelling. It is for these reasons, why system scientists should play an important role in the research required to attack the interdisciplinary problems in the modern life sciences.

3.4.2 Art Critics: Discussion

The role of mathematics and computation in describing, understanding, and modifying biological complexity has, in recent years, focused on bioinformatics and metabolic engineering. The area of bioinformatics, primarily occupied with pattern recognition and data mining, provides the biologist with data, information and facts from which he builds a model for inference in his mind. Metabolic engineering is the manipulation of industrial organisms using genetic tools. By identifying genes that confer a particular biochemical response or phenotype it has been related to drug discovery and functional genomics. So why should be look for novel mathematical models that go beyond those in metabolic engineering and biochemical analysis[19]?

[18] The description frame does constitute a system according to the definition given in Section 1.1 in that relations between objects are established by factors (induced equivalence relations) as well as concepts viewed as abstract objects for which (fuzzy) relations are established and considered in terms of approximate reasoning.

[19] The use of metabolic engineering as a methodology in biotechnology, control engineering and biochemical process analysis is not questioned here. The question is to what extend these models, in future, can help biologists solve basic biological (conceptual) problems (for instance in functional genomics)?

- ▷ Reductionism, modelling cell functions as a mechanism is bound to be limited to specific aspects of molecular systems in order for (linear) models to be tractable and to have explanatory value.
- ▷ Common assumptions and simplifications are that particular subsystems (e.g. a single gene's product) have a significant effect on the biochemical network.
- ▷ Current mathematical approaches are unable to capture the inherent complexity of biological systems which is due to interconnected and multileveled processes: Any pertubation of the cell will result in a multigene-multitranscript-multiprotein response but changes in one level do not neccessarily imply a corresponding change on other levels.

To consider an organism merely as a chemical machine is to take a reductionist approach in which the organisms's energy handling is explained in thermodynamic and chemical kinetic terms. Physics is relevant to biology but it can only deal with a part of the multi-levelled and highly interconnected causal hierarchies of biological systems. If the analysis of biological systems in terms of elements relating to the acquisition, transfer and utilisation of energy is not sufficient, what other component could complement the ergonic component? It is the study of everything functioning in the detection, processing, retention and utilisation of *information*[20]. As we have structured our presentation in analogy to the process of taking a photo, we notice that research has focussed on the analysis of images in terms of 'pixels', i.e., the detection of pattern. Though this provides *statistical information*, induced by its *elements*, it does not describe *conceptual information* about the *content* of the picture taken. In short, there is an important distinction between 'order' and the informed 'functional organisation'. Physical modelling and pattern recognition will therefore have to be integrated into a more 'holistic' framework if we are hoping to convince biologists of the usefulness of formal mathematical models.

In analogy to the question of ontology versus epistemology in philosophy, in science we can take alternative perspectives to modelling. Studying a biological phenomena, we may use systems of equations to model the chemical or biophysical structure of the process 'itself'[21]. Instead of attempting an ontological model of the thing itself, we may build a phenomenological model de-

[20]For a survey and discussion of the role of energy and information in modelling biological processes see for example [13].

[21]The basis for such thinking is that reductionism works, i.e., by decomposing a system into parts we can establish knowledge of the whole. Historically, this view and its consequences when generalised has led to a lack of respect for the complexity of nature. Marvel or fascination usually leads to a form of 'respect' while reductionist radicalism seems to have had more destructive consequences.

scribing "what we can know about it", i.e., representing measurements (samples taken experimentally) and observations[22] (facts or conclusions derived from perception or measurements) and constructing relationships that support the explanatory process. The role of uncertainty, the representation and quantification of uncertainty in measurements, observations and the model are of utmost importance in this approach.

Robert Rosen, arguing his case for relational biology, writes in [61] :

> "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."

One could argue, that up to now, biologists have succeeded without system theory and it may be doubtful whether a formal mathematical approach (for genetic networks) will ever be useful in providing new insights rather than just *simulating* what is already known. There are no theories of molecular biology like there are for instance in physics or statistics[23]. There seems not much effort in this direction either, as most biologist are occupied with extensive experimentation providing vast collections of data to be analysed. For many biologists the main problem in the progress of their field therefore appears to be the need for Information Technology (IT), i.e., tools for storage and search in such databases. In general, this leads to the impression that Computer Science provides the most important additions to progress in molecular biology. It is important to realise that IT alone will not be sufficient. In order to cope with complexity, not only interfaces to data sources but also conceptual tools, i.e., ways of *system theoretic thinking* are necessary to organise, structure the biologists enquiries which are at the root the identification of relations. Despite the Internet, providing an increasingly comprehensive collection of genome information coupled with sophisticated search interfaces, there is a need for concepts that summarise and capture the structure and interactions of components and (nonlinear, hierarchical) systems in their entirety, in a way, empirical or experimental techniques cannot. System theory, which is defined

[22]Take for example the measurement of the gene expression level of a particular gene. An observation could be the annotation of that gene in a database. If the annotation classifies the gene to a particular functional class, this fact or observation would originally be determined by means of experiments.

[23]Debates in biology do not seem to be about which technique is appropriate to solve a given problem but are rather more concerned with general aspects such as the interpretation of general principles. An example is the debate between ultra-darwinists (R.Dawkins), sociobiologists (E.Wilson), taking a gene-phenotype relationship as the basis for life, and opponents (e.g S.Rose, I.Lewontin) who suggest that there is more to life than is determined by genes alone.

as the study of organisation *per se*, may therefore provide not just *tools* but also a *way of thinking*. In this respect, system theory[24] may be a useful complement to the biologist's experimental work helping him to prevent pitfalls provided by the enemies of complexity and nonlinearity. I therefore believe, that the challenge in modern biology is an intelligent combination of human context-dependent expert knowledge, an interface to databases, a conceptual (system theoretic) framework as well as working methodologies in the analysis of data and simulation of systems.

3.5 FORMALITIES

Having claimed that basic problems in biology are conceptual and that these could be studied with abstract mathematical models, we ought to reconcile a conceptual framework with empirical research. In other words, to make our approach acceptable to experimental biologists, a model should be verifiable or identifiable with data. In the present section, we review our factor-space model in terms of equivalence relations induced by factors. This is going to open new pathways, establishing formal relationships to evidence theory and rough set theory, the latter of which has been used in data mining.

In the previous section, we noted that if we could integrate the factor-space approach into Rosen's relational theory of systems, we would have a powerful set of inferential tools available such as a 'language of models'. Such constructions would not only provide insights into biological phenomena, but we could also say something about the process of modelling itself – its limits, how much, how well we can infer causal entailment in a natural system from a formal model. This section therefore investigates the relationship between factor-space theory and Rosen's relational biology. Because a detailed overview of Rosen's work is beyond the scope of the present text, we will confine our discussion to pointing out the similarities of factors to *observables* and introduce the concept of *linkage* between observables in the context of our fuzzy relational factor-space model. In Section 3.5.1, we will demonstrate that if we accept the modelling relation and factors (observables) as an appropriate way of modelling biological phenomena, then uncertainty is certain. In Section 3.5.4, we establish a formal relationship to evidence theory which could serve as a means to capture probabilistic uncertainty if necessary. Finally, Rosen's concept of linkage, Shafer's inner/outer entailments and Pawlak's rough sets are shown to be closely related. With this package of connected research areas and tools established by them we hope for success in applying the fuzzy relational factor-space model to biological databases.

[24]System theory, like cybernetics, is a way of thinking, not a collection of facts. Thinking involves concepts: forming them and relating them to each other [20].

One way to describe a theory of systems is to define the three fundamental concepts *system*, *state*, and *observable*. An observable[25] of a system is some characteristic of it which can, in principle, be measured. Its formal definition is a mapping from a set of possible states to real numbers. Using the real line $\mathbb{R}$ as the domain has dominated system theory because of its mathematical structure and the assertion that natural processes are 'continuous'. Our definition of a factor has been somewhat more general but otherwise very similar to an observable, allowing us to borrow some ideas from Rosen's work, in particular the discussion of factors in terms of equivalence relations and the partitions they induce on U.

On page 55, we defined the family of factors $F \subset V$ available to us as *sufficient*, (3.4), such that for any two objects in U there exists a factor able to distinguish between the objects, i.e., the values of the factors on the objects differ, $f(u) \neq f(u')$. In many practical situations we cannot choose the objects and neither select an arbitrary way of measuring or observation. Given an abstract set U, representing the set of objects of some system, a factor defines a mapping $f\colon U \to X(f)$ but then also induces an *equivalence relation* E_f on U :

$$\begin{aligned} E_f\colon \quad U \times U &\to \{0,1\} \\ (u,u') &\mapsto E(u,u') = \begin{cases} 1 & \text{if } f(u) = f(u'), \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

A factor therefore forms *equivalence classes* from those objects $u \in U$ for which the factor f assumes the same value. That is, if $x = f(u)$, the class of u is the set $f^{-1}(u)$. Our choice for a factor may therefore lead to the situation in which two objects are indistinguishable by our measurement or observation. The imprecision or resolution of our factor or model in general is an important aspect for the description of biological systems and we shall discuss the issue in terms of equivalence classes following closely Rosen's work[26] [59].

Definition 11 (Equivalence Relations). Let U be a set, then a (crisp) relation R on U will be a subset of the Cartesian product, $R \subset U \times U$. If $(u,u') \in R$, we shall write $R(u,u') = 1$ (or simply $R(u,u')$ for short) to state that u is *related* to u' via R. A relation R in U is an *equivalence relation*,

[25] The term *observable* is well established not only in system theory but also quantum physics [5].
[26] In Rosen's definition of an observable, U is referred to as the set of *states* opposed to $X(f)$ in the factor-space model.

denoted E if is satisfies the following conditions

$$\begin{aligned} E(u,u) &= 1 && \forall\ u \in U && \text{(reflexive)},\\ E(u,u') &= 1 \ \Rightarrow\ E(u',u) = 1 && && \text{(symmetry)},\\ E(u,u') = 1 \ \wedge\ E(u',u'') &= 1 \ \Rightarrow\ E(u,u'') = 1 && && \text{(transitivity)}\ . \end{aligned}$$

Intuitively an equivalence relation is a generalisation of equality (which itself is an equivalence relation).

Definition 12 (Equivalence Class, Quotient Set). If E is an equivalence relation on U, and if $u \in U$, then the set of all objects u' such that $E(u,u')$ is the *equivalence class* of u under E, denoted $[u]_E$:

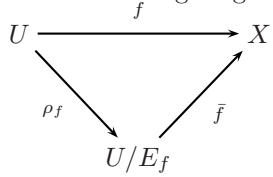
$$[u]_E = \{u' \in U\ :\ E(u,u') = 1\}\ .$$

By definition for every pair of objects u, u' in U we have $[u] = [u']$ or $[u] \cap [u'] = \emptyset$. That is, every object in U belongs to one and only one equivalence class under E. As a result of which U is effectively decomposed into subsets forming a *partition* of U. The set of equivalence classes of U under E is called the *quotient set* of U under E, denoted U/E :

$$U/E = \left\{[u]_E\ :\ u \in U \text{ and } [u]_E \cap [u']_E = \emptyset\right\}\ .$$

Let $f\colon U \to X$ be any mapping, and let E_f be the associated equivalence relation on U. Then there exists a one-to-one correspondence between $f(U)$, the range (also called *spectrum*) of f in X, and U/E_f. In other words, for any $x \in f(U)$, there exists an $u \in U$ such that $f(u) = x$ where each x can be identified with the equivalence class $[u]_{E_f}$ (also denoted $[u]_f$ for short). Similar, an important fact is that there exists a one-to-one correspondence between the equivalence relations on a set U and the partitions of U which effectively allows us to shift our discussion about factors on U to equivalence classes or equivalently quotient sets on U. These ideas effectively establish the relationship of the fuzzy relational factor-space model with relational biology. The formal link to fuzzy mathematics is introduced in the next section while in subsequent sections we will discuss modelling of systems using factors. The choice of factors and their effectiveness in describing the process under consideration is then established by means of the equivalence relations and partitions they induce on U.

Denoting the one-to-one correspondence, between $f(U)$ and U/E_f, by $\bar{f}$, it is the mapping that makes the following diagram commutative.



Where $f(U)$ is a subset of X and the mapping

$$\begin{aligned} \rho\colon \quad U &\to U/E_f \\ u &\mapsto \rho(u) = [u]_{E_f} \end{aligned}$$

is called the *natural mapping* of U onto U/E_f. Hence, using factor f, what we observe is not the set of objects, but rather the quotient set U/E_f. The set U/E_f is therefore referred to as by Rosen as the set of *reduced states* of U under f and for any given $u \in U$, the corresponding equivalence class $[u]_{E_f}$ is called a *reduced state* of u. The space of reduced states plays then the role of the 'state-space' or 'phase-space' in control theory and physics respectively. The special case in which $X \equiv \mathbb{R}$ implies that U/E_f is a topological space and hence allows an analysis of any two $f(u)$ in terms of distances or a metric between them. The books by Rosen [60, 59] provide a rich source of material for this case which is closely related to Von Neumann's approach to quantum physics [5].

3.5.1 Through the Blurred Looking Glass

As suggested in Figure 2.9, if we are to reduce the process of a scientific investigation to two concepts, it would be *comparing* and *reasoning*. We use sets and operations on sets or, equivalently, relations in order to group and hence compare objects. In the factor-space model, we use equivalence relations as a means to validate the effectiveness of a factor – our tool to probe or describe a biological system. A factor is, in general, designated by a noun, a state by a numeral, and a characteristic by an adjective. For example, gene expression may be a factor; the amount/level of a protein measurable is referred to as a state and the judgement of a "high level" is a characteristic. Types of factors may be distinguished by their state-space $X(f)$. For measurable factors such as time, length, mass, etc. we usually take the real line $\mathbb{R}$ or subsets while for degrees we may take the unit interval $[0, 1]$ as the state-space. The elements x of $X(f)$ may however also be more qualitative such as switching factors whose values may for instance come from the set {"yes", "no"}. This section will discuss problems that may occur when using equivalence relations in an experimental context where imprecision in form of measurement errors in unavoidable.

The previous section introduced equivalence relations for which equality ($=$), and elementhood ($\in$) are examples. With respect to equality, it is obvious that it satisfies the three conditions for an equivalence relation (Definition 11) :

$$\begin{aligned} a = a \quad &\text{holds} && \text{(reflexity)} \\ a = b \;\Rightarrow\; b = a \quad & && \text{(symmetry)} \\ a = b \;\wedge\; b = c \;\Rightarrow\; a = c \quad & && \text{(transitivity)}\ . \end{aligned}$$

Transitivity enables us to infer something new about the relationship of two variables given two pieces of information. The concept plays consequently a fundamental role in reasoning. We can illustrate transitivity with another important tool for comparison: the concept of a distance or metric.

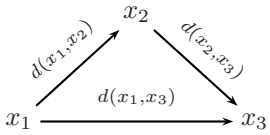
Definition 13 (Distance). The function $d(\cdot,\cdot)$ defines a *distance* between elements of X. Let for any x_1, x_2, x_3 in X :

$$\begin{aligned} &d(x_1,x_2)=0 && \text{iff} \quad x_1=x_2 \\ &d(x_1,x_2)>0 && \text{iff} \quad x_1\neq x_2 \\ &d(x_1,x_2)=d(x_2,x_1) && \text{symmetry .} \end{aligned}$$

Definition 14 (Metric). A distance is called *metric* iff $\forall x_1, x_2, x_3 \in X$ it is transitive :

$$d(x_1,x_3) \leq d(x_1,x_2)+d(x_2,x_3) \tag{3.38}$$

called *triangle inequality* :



A simple example for a metric is the absolute value of the difference

$$d(x,x') = |x-x'| \ .$$

Having established the basic tools for comparison and reasoning it remains to define a mechanism of *order*. Examples of relations which establish an order are the 'greater than' and 'subsethood' relations. Formally, they are establishing a *partial order* on X, making X a partially ordered set (poset). Again these relations are transitive :

$$\begin{aligned} &\text{"greater or equal"} \quad \geq : \quad x_1>x_2 \quad \wedge \quad x_2>x_3 \quad \Rightarrow \quad x_1>x_3 \\ &\text{"set inclusion"} \qquad \subseteq : \quad A\subseteq B \quad \wedge \quad B\subseteq C \quad \Rightarrow \quad A\subseteq C \end{aligned}$$

Definition 15 (Partial Order). A *partial ordering* (or semi-ordering) on X is a binary relation $\preceq$ on X such that the relation is

reflexive, i.e., $x \preceq x$,
anti-symmetric, i.e., $x \preceq x'$ and $x' \preceq x$ implies $x = x'$,
transitive, i.e., $x \preceq x'$ and $x' \preceq x''$ implies $x \preceq x''$.

In our previously established framework we are required to establish the equality of states, i.e., the values of a factor on an object u in U, $f(u) = f(u')$. Accepting some imprecision in the measurement of values, we face a major problem matching our theory with physical reality. The problem, referred to as the Poincaré paradox describes the *indistinguishability* of individual elements in non-mathematical continua. More specifically, for three points x_1, x_2, x_3 and let ε denote a threshold (tolerance, significance level), then two elements x, x' are indistinguishable for $d(x, x') \leq \varepsilon$, where $d(\cdot, \cdot)$ denotes a proximity measure such as a metric defined above. Transitivity for (pseudo)metrics manifests itself in the *triangle inequality*

$$d(x_1, x_3) \leq d(x_1, x_2) + d(x_2, x_3) \ . \tag{3.38}$$

Let us consider the following measurements in $\mathbb{R}$: $x_1 = 1.5,\ x_2 = 2,\ x_3 = 2.2,$ and $\varepsilon = 0.6$, and let us use the metric $d(x, x') = |x - x'|$, w.r.t a threshold, accuracy or error bound ε, to identify observations. Our analysis is based on the *theoretical model* that if observations x_1 and x_2 are similar, as well as x_2 and x_3 are similar, then so should be x_1 and x_3, in other words,

$$x_1 = x_2 \ \wedge \ x_2 = x_3 \ \Rightarrow \ x_1 = x_3 \ .$$

Now, considering measured data,

$$\begin{aligned} & |x_1 - x_2| = 0.5 \ < \varepsilon \quad \Rightarrow \quad x_1 = x_2 \\ & |x_2 - x_3| = 0.2 \ < \varepsilon \quad \Rightarrow \quad x_2 = x_3 \\ \text{but} \quad & |x_1 - x_3| = 0.7 \ > \varepsilon \quad \Rightarrow \quad x_1 \neq x_3 \ . \end{aligned} \tag{3.39}$$

Fortunately, a solution to this dilemma (uncertainty is certain!) leads us directly to fuzzy sets which we have incorporated in our factor-space model right from the start. To bridge the mathematical idealisation with physical reality, Karl Menger suggested a measure between 0 and 1, probabilities, to quantify uncertainty while retaining the all important property of transitivity. Specifically, he associated $d(x, x')$ with a (cumulative) distribution function $F_{x,x'}$ whose value $F_{x,x'}(a)$ for any a is interpreted as the probability that the distance between x and x' is less than a. Menger's approach generalises a metric space to become a *probabilistic metric space* [68].

Starting with the triangle inequality (3.38) we note that it implies the logical proposition

$$d(x_1, x_2) < a \quad \wedge \quad d(x_2, x_3) < b \ \Rightarrow \ d(x_1, x_3) < a + b \ .$$

Since $A \Rightarrow B$ implies that $Pr(A) \leq Pr(B)$, we get

$$Pr\big(d(x_1, x_2) < a \ \wedge \ d(x_2, x_3) < b\big) \leq Pr\big(d(x_1, x_3) < a + b\big) = F_{x_1 x_3}(a + b)$$

Thus if T is such that $T(Pr(A), Pr(B)) \leq Pr(A \wedge B)$ for any two propositions A, B, we obtain the inequality

$$F_{x_1 x_3}(a + b) \geq T\big(F_{x_1 x_2}(a), F_{x_2 x_3}(b)\big) \ , \tag{3.40}$$

where T denotes a triangular norm, so called because it generalises the triangle inequality. The function T is a mapping $[0,1] \times [0,1] \to [0,1]$. For example,

$$\begin{aligned} T_{\min}(a,b) &= \min(a,b) && \text{(minimum operator)}, \\ T_{\text{Luk}}(a,b) &= \max(a+b-1,0) && \text{(Lukasiewicz norm)}, \\ T_{\text{pro}}(a,b) &= a \cdot b && \text{(algebraic product)}. \end{aligned} \tag{3.41}$$

Let $T = a \cdot b$, then (3.40)

$$F_{x_1x_3}(a+b) \geq F_{x_1x_2}(a) \cdot F_{x_2x_3}(b) \tag{3.42}$$

states that the probability of the distance between x_1 and x_3 being smaller than $a+b$ is at least the joint probability of the independent occurrence of the distance between x_1 and x_2 being smaller than a and the distance between x_2 and x_3 being smaller than b. In other words,

$$Pr\big(d(x_1,x_3) < a+b\big) \geq Pr\big(d(x_1,x_2) < a,\ d(x_2,x_3) < b\big)\ .$$

So much for probabilistic uncertainty and the introduction to triangular norms. We now show that a metric induces a similarity (fuzzy) relation for which transitivity is generalised in the form of inequality (3.40). In the factor-space model, values x in $X(f)$, referred to as states, are in fact evaluations of factors on objects in U. That is, for two such values being very close, $f(u) \approx f(u')$, two objects u and u' are indistinguishable by f, their values are similar and hence, with respect to f, they are equivalent. This observation motivates the definition of a *fuzzy equivalence relation*, $\widetilde{E}$ as a direct generalisation of the crisp equivalence relation E in Definition 11.

Definition 16 (Fuzzy Equivalence Relations). A *fuzzy equivalence* or *similarity relation*, $\widetilde{E}$, is a fuzzy relation which is reflexive, symmetric, and transitive. It defines a function $\widetilde{E}\colon U \times U \to [0,1]$ that satisfies the conditions :

$$\begin{aligned} \widetilde{E}(u,u) &= 1 \qquad \forall\ u \in U && \text{(reflexive)}, \\ \widetilde{E}(u,u') &= \widetilde{E}(u',u) && \text{(symmetric)}, \\ \widetilde{E}(u,u'') &\geq T\big(\widetilde{E}(u,u'), \widetilde{E}(u',u'')\big) && \text{(transitive)}\ . \end{aligned}$$

Transitivity for fuzzy relations is therefore defined in analogy to Menger's inequality, (3.40), for probabilistic metric spaces. In this context, the triangular norm, T, extends the domain of logical conjunction from the set $\{0,1\}$ to the interval $[0,1]$. Using the min-operator, we speak of min-transitivity as a natural extension of the equivalence relation above. The equivalence classes partition U into sets containing elements that are all similar to each other to degree at least ε.

For a bounded metric space (X, d) there exists a non-negative value $\varepsilon \in \mathbb{R}^+$ such that $d(x, x') \leq \varepsilon$, for all x in X. The distance d between values of factors on objects then induces a fuzzy relation over U :

$$\widetilde{E}(u, u') = 1 - \frac{1}{\varepsilon} d\big(f(u), f(u')\big) \ . \tag{3.43}$$

The bound ε allows scaling such that the distance between any two values in X lies in the unit interval $[0, 1]$. The correspondence of transitivity for a distance function and transitivity of fuzzy relations depends on the T-norm employed. The metric equivalent of the Lukasiewicz norm, (3.41), is the triangle inequality, (3.38); the metric equivalent of product transitivity is the inequality

$$d(x_1, x_3) \leq d(x_1, x_2) + d(x_2, x_3) - d(x_1, x_2) d(x_2, x_3)$$

related to (3.42) w.r.t probabilistic uncertainty. If a fuzzy equivalence relation is min-transitive the distance satisfies the more restrictive ultrametric inequality :

$$d(x_1, x_3) \leq \max\big(d(x_1, x_2), d(x_2, x_3)\big) \ .$$

The Lukasiewicz norm turns out to be the least restrictive one. For the comparison of factors on objects u it would usually be reasonable to assume that to objects are similar in their contribution to the model if $|f(u) - f(u')| \leq \varepsilon$, where ε is a number representing our "indifference" with respect to the measurement process.

We saw that if we are to use equivalence relations, in practical situations, we may allow for a tolerance to identify two objects as the same (as having the same observable consequence). The inequality $|f(u) - f(u')| \leq \varepsilon$ describes a subset (relation) $R_\varepsilon \subset U \times U$

$$R_\varepsilon = \{(u, u') \in U \times U \ : \ |f(u) - f(u')| \leq \varepsilon\} \ .$$

The Poincaré paradox (3.39) demonstrated that this relation is not an equivalence relation, i.e., it is not a transitive relation. We therefore could not study the quotient set induced by this relation. Kruse et al. [32][27] showed however that we can define a mapping $\widetilde{E}_\varepsilon$ such that $\widetilde{E}_\varepsilon(u, u')$ is greater than $1 - \varepsilon$ if and only if u and u' are indistinguishable with respect to the tolerance ε :

$$(u, u') \in R_\varepsilon \quad \text{if and only if} \quad \widetilde{E}_\varepsilon(u, u') \geq 1 - \varepsilon \ ,$$

[27] Their book [32] and various papers by R.Kruse and F.Klawonn provide an extensive treatment of equivalence relations and how rule-based systems can be build from them. They also generalise the case described here with $\varepsilon \in [0, 1]$ and $\widetilde{E}_\varepsilon(u, u') = 1 - \min\{|f(u) - f(u')|, 1\}$ to any unit in X by means of a scaling factor $s > 0$, $\widetilde{E}_\varepsilon(u, u') = 1 - \min\{|s \cdot f(u) - s \cdot f(u')|, 1\}$.

where

$$\begin{aligned} \widetilde{E}_\varepsilon : \quad U \times U &\to [0,1] \\ (u,u') &\mapsto 1 - \inf\{\varepsilon \in [0,1] \,:\, (u,u') \in R_\varepsilon\} \end{aligned}$$

with $\varepsilon \in [0,1]$ and if there is no ε for which the relation holds, we define $\inf \emptyset \doteq 1$. $\widetilde{E}_\varepsilon$ is a fuzzy equivalence relation w.r.t. T_{Luk}. The value $\widetilde{E}_\varepsilon(u,u') = 1 - \min\{|f(u) - f(u')|, 1\}$ describes the degree to which two objects u and u' have similar observable consequences and transitivity of this relation implies that if u and u' are similar and u' and u'' are similar in their values in X, then u is similar to u''.

Remark. Similarity relations were introduced by Zadeh [82] and have since been considered in various contexts. For example in clustering, fuzzy equivalence classes $[u]_{\widetilde{E}}$ define clusters while the fuzzy quotients $U/\widetilde{E}$ partition U. The mapping from elements of U to equivalence classes in $U/\widetilde{E}$ then defines a classifier. The formal setting of fuzzy equivalence relations, classes and quotients in fuzzy mathematics and category theory has been developed largely by Höhle [22, 23, 24].

At the beginning of this section, we introduced equivalence relations as a fundamental notion for comparing, ordering and reasoning. We showed that in an experimental context the relation $|f(u) - f(u')| \leq \varepsilon$ is reflexive, symmetric but not transitive on $\mathbb{R}$. As a consequence, uncertainty in analysis is certain and a extension of the classical definition of equivalence, first suggested by Karl Menger, seems sensible. The following section explores the use of (classical) equivalence relations to assess the role of factors in modelling process itself.

3.5.2 The Art of Modelling: Linkage

In molecular biology, the objects which constitute a system are usually not directly accessible for measurement. We therefore introduced factors in order to describe a *concept* by means of observable characteristics. Following the definition of a factor as a mapping in Section 3.1.3 and the discussion of its properties in terms of equivalence classes, we are now in the position to discuss or compare 'different' ways to describe the same process. Let us therefore suppose we are given two factors $f, g \in F$ such that for each $u \in U$ we have two 'coordinates', $f(u)$ in U/E_f and $g(u)$ in U/E_g, as independent descriptions of the same concept. Our discussion extends the definitions 3, 4, 5 but now with the focus on how to improve our description of a concept through equivalence classes induced by the factors chosen. In Section 3.5.4, we are going to introduce yet another way to discuss the relationship between factors. A measure of linkage between factors f and g, is devised based on

equivalence relations and can be used to identify cause-effect relationships in data.

What follows is an introduction of Rosen's concept of *linkage* between factors. We shall discuss three cases for which two factors are 'unlinked', 'linked' and 'partially linked'. First consider the illustration in Figure 3.13 defining two factors f and g which partition U in different ways.

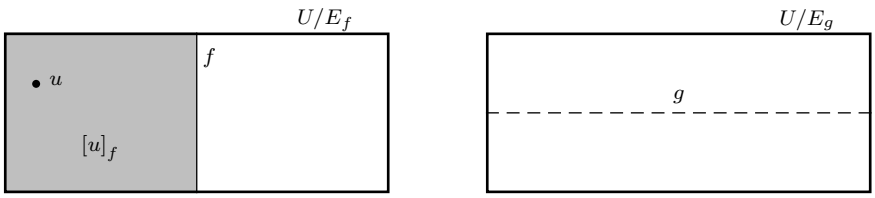


Fig. 3.13 Example of two totally unlinked factors f and g. They grey area on the left is the equivalence class $[u]_f$ generated by f on U.

The concept of linkage between factors f and g becomes plausible by assuming a given $[u]_f$ in U/E_f and subsequently to discuss which g-equivalence classes intersect with $[u]_f$. From Figure 3.13, we find that factor g splits the classes of E_f such that g can distinguish between objects, undistinguishable via f. We say that the greater the extend of the splitting of $[u]_f$ by g, the more unlinked g is to f at $[u]_f$. We find that

- The whole of U/E_g, i.e., both g-classes intersect with $[u]_f$: g is said to be *unlinked* to f at $[u]_f$.
- g is unlinked to f at each $[u]_f$; every E_f-class intersects every E_g-class and conversely : g is said to be *totally unlinked* to f.

Having fixed some value x in $f(U)$, $g(u)$ is *not* arbitrary in $g(U)$; the coordinates $f(u)$, $g(u)$ of an object $u \in U$ are not independently variable in U/E_f, U/E_g, respectively.

Figure 3.14 illustrates the second extreme: total linkage. We make the following observations :

- Only a single g-class intersects with $[u]_f$: g is said to be *linked* to f at $[u]_f$.
- Since g is linked to f at each $[u]_f$; every class of E_f intersects exactly one class of E_g, namely the one which contains it : g is said to be *totally linked* to f.

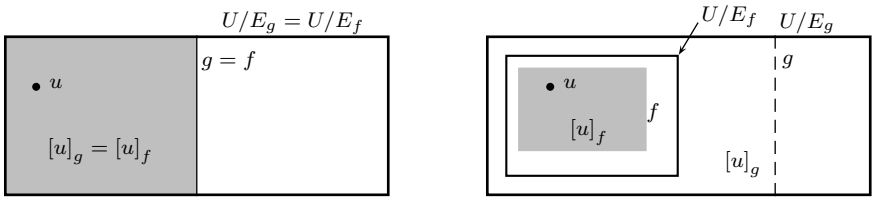


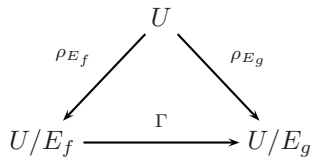
Fig. 3.14 Two examples of two totally linked factors f and g such that E_f refines E_g.

If g and f are totally linked, E_f is said to *refine* E_g, g does not split the classes of E_f and no new information is obtained from an additional factor g. The coordinates $f(u)$ and $g(u)$ are *independently variable* in U/E_f, U/E_g respectively. That is, having fixed some value x in $f(U)$ we may find an object in U such that $f(u) = x$ and $g(u)$ is arbitrary in $g(U)$.

In general, let E_f, E_g be equivalence relations on a set U. E_f is said to be a refinement of E_g if $E_f(u, u')$ implies $E_g(u, u')$. In terms of equivalence class, this means that every E_f-equivalence class is a subset of some E_g-equivalence class or in other words, E_f refining E_g means that elements of the partition from E_g are further partitioned by E_f and blocks of the E_g partition can be obtained from the set-theoretic union from E_f-blocks. If E_f is a refinement of E_g, then there is a unique mapping

$$\begin{aligned} \Gamma\colon \quad U/E_f &\rightarrow U/E_g \\ [u]_f &\mapsto \Gamma([u]_f) = [u]_g \end{aligned} \tag{3.44}$$

which makes the following diagram commute :



Thus the value of g on an object u in U is completely determined by the value of f on that object through the relation $g(u) = \Gamma(f(u))$. That is, g is a function of f. Next, let $f, g\colon U \rightarrow \{0, 1\}$ be defined such that its value is equal to one if u is on the right of the line which partitions U and zero otherwise. We then have the situation depicted on the right in Figure 3.15 where find that :

- For u_1, only one g-class intersects with $[u]_f$ but not all of U/E_g. That is, g is linked to f at $[u]_f$.
- For u_2, both g-classes intersect with $[u]_f$ and hence g is unlinked to f at $[u]_f$.

We also note that the linkage relationship between f and g is not symmetric; i.e., the linkage of g to f at $[u]_f$ can be different from the linkage of f to g. Here we have motivated the concept of linkage as a means to discuss the usefulness of additional factors in our model while in Rosen's work its importance comes from the possibility of prediction. That is, if g is linked to f at u, we can determine information about $g(u)$ via f.

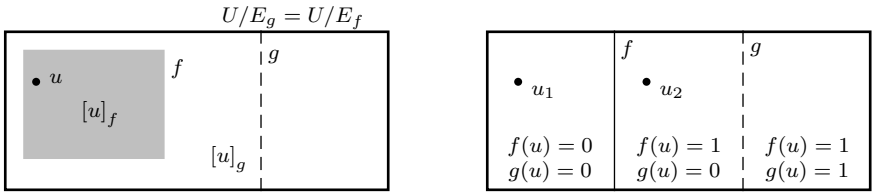


Fig. 3.15 Two examples for partial linkage between factors.

Before concluding this subsection, we look at another illustration of linkage. From Figure 3.16, we have the following equivalence classes for f and g from which we find that f and g are totally unlinked.

$$
\begin{array}{lll}
[u_1]_f = \{u_1, u_2\} & [u_1]_g = \{u_1, u_3\} & U/E_f = \big\{\{u_1, u_2\}, \{u_3, u_4\}\big\} \\
[u_2]_f = \{u_1, u_2\} & [u_2]_g = \{u_2, u_4\} & U/E_g = \big\{\{u_1, u_3\}, \{u_2, u_4\}\big\} \\
[u_3]_f = \{u_3, u_4\} & [u_3]_g = \{u_1, u_3\} & \\
[u_4]_f = \{u_3, u_4\} & [u_4]_g = \{u_2, u_4\} & U/E_{fg} = \big\{\{u_1\}, \{u_2\}, \{u_3\}, \{u_4\}\big\}
\end{array}
$$

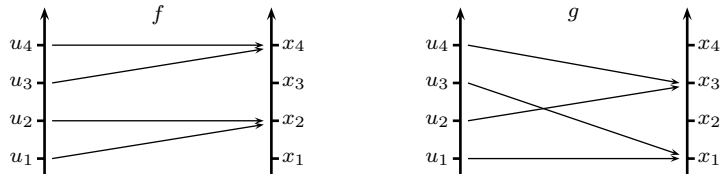


Fig. 3.16 Example of two totally unlinked factors f and g.

In Figure 3.17, we find an example for total linkage. The equivalence classes and quotient sets are as follows.

$$
\begin{array}{lll}
[u_1]_f = \{u_1, u_2\} & [u_1]_g = \{u_1\} & U/E_f = \big\{\{u_4\}, \{u_3\}, \{u_1, u_2\}\big\} \\
[u_2]_f = \{u_1, u_2\} & [u_2]_g = \{u_2\} & U/E_g = \big\{\{u_3, u_4\}, \{u_2\}, \{u_1\}\big\} \\
[u_3]_f = \{u_3\} & [u_3]_g = \{u_3, u_4\} & \\
[u_4]_f = \{u_4\} & [u_4]_g = \{u_3, u_4\} & U/E_{fg} = \big\{\{u_1\}, \{u_2\}, \{u_3\}, \{u_4\}\big\}
\end{array}
$$

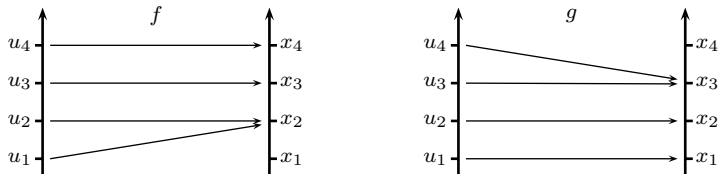


Fig. 3.17 Example of two totally linked factors f and g.

Finally we look at an example for partial linkage, illustrated in Figure 3.18. The equivalence classes and quotient sets are :

$$
\begin{array}{lll}
[u_1]_f = \{u_1, u_2\} & [u_1]_g = \{u_1\} & U/E_f = \big\{\{u_1, u_2\}, \{u_3, u_4\}\big\} \\
[u_2]_f = \{u_1, u_2\} & [u_2]_g = \{u_2, u_3\} & U/E_g = \big\{\{u_1\}, \{u_2, u_3\}, \{u_4\}\big\} \\
[u_3]_f = \{u_3, u_4\} & [u_3]_g = \{u_2, u_3\} & \\
[u_4]_f = \{u_3, u_4\} & [u_4]_g = \{u_4\} & U/E_{fg} = \big\{\{u_1\}, \{u_2\}, \{u_3\}, \{u_4\}\big\}
\end{array}
$$

With respect to the linkage of g to f we find that for all u in U, g is *partially linked* to f at $[u]_f$ since it intersects with more than one g-class but not all of U/E_g. The linkage of f to g at $[u]_g$ is however different :

- Linkage at $[u_1]_g$: Intersects with a single f-class.
- Unlinked at $[u_2]_g$ and $[u_3]_g$: Intersections with all of U/E_f.
- Linkage at $[u_4]_g$.

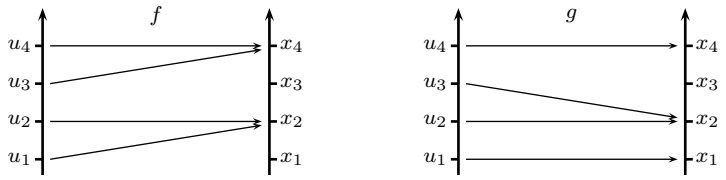


Fig. 3.18 Example of partial linkage between f and g.

3.5.3 The Art of Modelling: Product Space Representation

In this section, we show how a family of independent atomic factors defined on U can be used to obtain a more comprehensive description of the elements of U than is possible with one factor. The following therefore extends equation (3.3) of the Cartesian product of independent factors. The aim of the product

representation is to obtain for a given $u \in U$ a unique representation in form of 'coordinates' :

$$\begin{aligned} U &\rightarrow X(f) \times X(g) \\ u &\mapsto \big(f(u), g(u)\big) \ . \end{aligned}$$

In order to improve our description of the system via U, the equivalence relation of both factors, E_{fg} should lead to a finer partition than from either f or g alone. Taking the intersection, $E_{fg} = E_f \cap E_g$, such that $E_{fg}(u, u')$ iff $E_f(u, u')$ and $E_g(u, u')$, we find the equivalence classes as $[u]_{fg} = [u]_f \cap [u]_g$. The new equivalence relation does only hold if and only if $f(u) = f(u')$ and $g(u) = g(u')$. We thus seek an embedding

$$\phi\colon\ U/E_{fg} \ \rightarrow\ U/E_f \times U/E_g \ . \tag{3.45}$$

Rosen [59] derived the following conditions for ϕ :

- The embedding ϕ is onto if and only if f and g are totally unlinked.
- If ϕ is not onto, there is linkage between f and g, that is, not all pairs in $U/E_f \times U/E_g$ represent objects in U (represent equivalence classes of objects under E_{fg}).
- If E_f refines E_g, U/E_{fg} is a 'curve' in $U/E_f \times U/E_g$; i.e., a one-dimensional subspace.
- If $E_f = E_g$, the projections of this curve on each of the 'coordinate axis' U/E_f and U/E_g is one-one onto; the curve is the graph of the function h (3.44).

The case for two factors can be generalised to an arbitrary family of factors $G = \{f_1, f_2, \ldots, f_r\}$ corresponding to the definition of atomic factors on page 55 and the G-envelope (page 64). With $\prod$ denoting the Cartesian product, we define for the atomic factors $f_i \in G$ the one-to-one mapping

$$\phi\colon\ U/E_G \ \rightarrow\ \prod_{j=1}^{r} U/E_{f_j}$$

where each $[u]_G$ is the unique intersection of the classes $[u]_{f_j}$ for each f_j in G such that each $[u]_G$ is associated with the sequence of numbers $\big(f_j(u)\big)$ uniquely determined in $\prod_{f_j \in G} U/E_{f_j}$ and denoted by $\phi([u]_G)$. For the mapping ϕ to be onto the factors in G have to be pairwise unlinked.

Since in experimental biology, we may not be sure about the extend of the linkage between factors *a priori*, a measure of the magnitude between two factors f and g in an object u is desirable. From the preceding argument such a measure can be regarded as the number of distinct g-classes which intersect

$[u]_f$. Each of these distinct g-classes $[u']_g$ gives rise to a distinct element in the set U/E_{fg} and hence a distinct point $(f(u'), g(u'))$ in the Cartesian product $U/E_f \times U/E_g$. If f and g are totally linked, there is only one such point; if they are unlinked, then this set is of the form $\{[u]_f\} \times U/E_g$. Rosen therefore defines the magnitude of linkage by the number of points of the form $(f(u), g(u'))$ in $U/E_f \times U/E_g$ which are actually representatives of elements in U/E_{fg}. If the spectrum $g(U)$ is finite, the magnitude of the linkage of g to f at object u is defined as

$$\mathcal{L}(f,g) = \frac{\eta\big(\phi([u]_f) \cap [u']_g\big)}{\eta\big(g(u)\big)} \tag{3.46}$$

where u is fixed and u' is variable over U. Finally, another way of measuring the linkage between g and f can be based on replacing a given object u by a new object u', such that $f(u) - f(u')$ but $g(u) \neq g(u')$.

The key to our ability to understand the world around us is to form relations between percepts matching those between objects in the real world. In the formal approach presented here percepts are represented by abstract objects and/or concepts. The choice of factors to observe a system or the change to a system, observing its response, makes the scientific enquiry, captured by the modelling relation, an art.

3.5.4 Double Vision: Evidence Theory and Rough Sets

In the present section we are going to explore affinities of the factor-space approach to evidence theory (Dempster-Shafer) or rough set theory (Pawlak). We are interested in such link for two reasons. Evidence theory provides a way of allowing for partial (probabilistic) evidence in the description of concept C in U via its extension A. Secondly, rough set theory is said to have been successful in data mining applications and hence these results could be beneficial in implementing the fuzzy relational factor space approach and applying it to biological databases.

Looking first at evidence theory [69], U is referred to as a *frame of discernment*; a set of alternatives perceived as distinct answers to a question. Let $\mathcal{P}(U)$ denote the set of subsets of U (the power set). Whereas the degrees of membership $\widetilde{A}(u)$ are specifying the relevance of u to concept C, (cf. Section 3.1.2), partial evidence in terms of probabilities is modelled in evidence theory by considering a *mass distribution* (probability assignment) $m\colon \mathcal{P}(U) \to [0,1]$ where $m(\emptyset) = 0$ and $\sum_{A\colon A \subseteq U} m(A) = 1$. These are in fact axioms defining a probability measure. $m(A)$ is understood as a measure of belief committed to A. If $m(A)$ is not known exactly but partial evidence exists for subsets B of U, the following two real-valued functions describe the *belief* and *plausibility*

of A, respectively :

$$\begin{aligned} \text{Bel}: \quad \mathcal{P}(U) &\rightarrow [0,1] \\ A &\mapsto \text{Bel}(A) \doteq \sum_{B\,:\,B\subseteq A} m(B)\ . \end{aligned}$$

and

$$\begin{aligned} \text{Pl}: \quad \mathcal{P}(U) &\rightarrow [0,1] \\ A &\mapsto \text{Pl}(A) \doteq \sum_{B\,:\,B\cap A\neq\emptyset} m(B)\ . \end{aligned}$$

Whereas the book by Shafer [69] is a comprehensive introduction to the theory, the book by Kruse and Schwecke [31] provides an extensive treatment of evidence theory, its relation to possibility theory and their implementation in knowledge-based systems. The generalisation of belief functions to deal with fuzzy sets $\widetilde{A}$ is described in [81].

In his book, Shafer also discussed ways of comparing two frames of discernment and in particular how one frame can be obtained from another by refinement. We find that our discussion on linkage in Section 3.5.2 can be rephrased in evidence theory. What has been a discussion about additional factors is now the study of *frames that are different but compatible*. A frame being compatible means that it does not provide contradictory information but instead refines in some way the description of the concept of concern. What follows is a mathematical representation of how one frame of discernment U' is obtained from another frame of discernment U by splitting (refining) some or all of the elements of U. Following closely Shafer's description we introduce the mapping Γ which for each $u \in U$, defines a subset $\Gamma(\{u\})$ of U'. The sets $\Gamma(\{u\})$ are required to be non-empty, $\Gamma(\{u\}) \neq \emptyset$, and together form a *partition*, that is, the sets $\Gamma(\{u\})$ are disjoint, non-empty and their union form U'. The mapping (cf. (3.44))

$$\begin{aligned} \Gamma: \quad \mathcal{P}(U) &\rightarrow \mathcal{P}(U') \\ A &\mapsto \Gamma(A) = \bigcup_{u\in A} \Gamma(\{u\}) \end{aligned}$$

is called a *refining* and U' is said to be the refinement of U. Equivalently, U may be seen as the *coarsening* of U' as illustrated in Figure 3.19. In terms of two factors, f and g, we then have

$$\begin{aligned} \Gamma(\{u\}) &= \Gamma\big([u]_f\big) \\ &= [u]_g\ . \end{aligned}$$

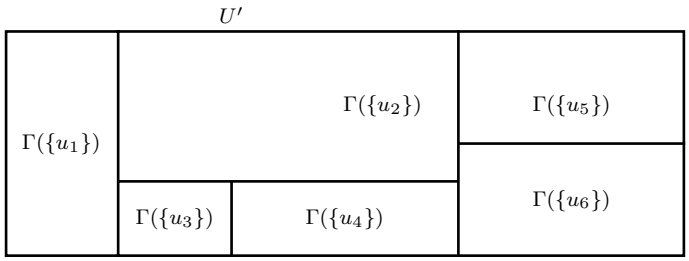


Fig. 3.19 A coarsening $U = \{u_1, \ldots, u_6\}$ of frame U' [69].

A frame of discernment, U, is understood as a set of alternative propositions perceived as distinct conclusions to a hypothesis. If the refinement Γ exists, U and U' are said to be *compatible*. The concept of refinement is a tool to compare two frames. On the other hand, coarsening is equivalent to clustering elements by building a partition on U. Therefore considering only one factor f, U/E_f is a coarsening of U, and U is a refinement of U/E_f. Then for $[u]_f \in U/E_f$, $\Gamma([u]_f)$ defines a subset of U and for any $B \subset U/E_f$,

$$\Gamma(B) = \bigcup_{[u]_f \in B} \Gamma\big([u]_f\big) \ .$$

In the context of comparing two compatible frames, associated with a refinement $\Gamma\colon \mathcal{P}(U) \to \mathcal{P}(U')$, Shafer also defines for a subset A of U the following two sets, called inner and outer reduction respectively :

$$R_* = \{u \in U \ : \ \Gamma(\{u\}) \subset A\},$$
$$R^* = \{u \in U \ : \ \Gamma(\{u\}) \cap A \neq \emptyset\} \ .$$

In our context, using factor f, inducing the equivalence relation E_f on U, we therefore define for a subset A of U, two subsets of U/E_f, the *inner reduction*

$$E_*(A) = \{[u]_f \ : \ [u]_f \subseteq A\} \tag{3.47}$$

and *outer reduction*

$$E^*(A) = \{[u]_f \ : \ [u]_f \cap A \neq \emptyset\} \ . \tag{3.48}$$

As illustrated in Figure 3.20, if we take A to be the grey shaded subset of U, then

$$E_*(A) = \{[u_3]_f\},$$
$$E^*(A) = \{[u_1]_f, [u_2]_f, [u_3]_f, [u_4]_f\} \ .$$

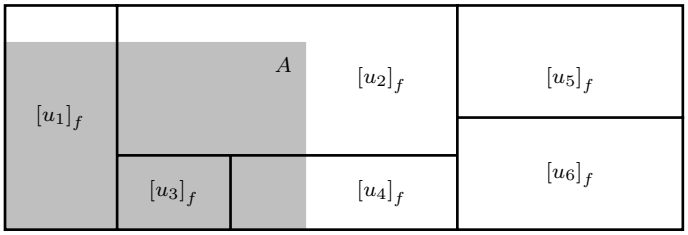


Fig. 3.20 Inner and outer reductions [69].

Reminding ourselves of the meaning of $\widetilde{A} \in \mathcal{F}(U)$ being the extension of a concept C in U (the biological phenomena in question), we find that for the crisp set $A \in \mathcal{P}(U)$ the pair E_*, E^* of subsets of U/E_f represents in fact an *approximation* of A from the outside and inside respectively. With the factor f being our only way of practically describing the concept C in terms of measurements, we can only observe the quotient set (coarsening) U/E_f of U w.r.t. E_f. Note also that an equivalence class $[u]_f$ consists of those elements u' of U for which $f(u) = f(u')$ and $f(u), f(u') \in X(f)$.

In Definition 7 on page 59, we introduced a measure of coincidence of the actual $\widetilde{A}$ with the representation extension $f(\widetilde{A})$ of C in $X(f)$ which we can observe via f. With E_* and E^* we have now a similar way to discuss the approximation of a (crisp) subset A. In [18], Dubois and Prade point out that the pair of subsets $E_*(A)$, $E^*(A)$ of U/E_f obtained from Shafer's inner/outer reductions, are in fact what Pawlak [54] later called a *rough set*. In rough set theory, $E^*(A)$ (resp. $E_*(A)$) are called *upper* (*lower*) *approximation* of A by E_f. $E_*(A) \subseteq E^*(A)$ and whenever $E^*(A) \neq E_*(A)$, $A \subset U$ cannot be perfectly described because of the *indiscernibility* of objects in U. Rough set theory has claimed success in data mining applications [55, 56, 83, 84] and it may therefore be useful to familiarise ourselves with its terminology in order to investigate its application to biological databases.

In rough set theory, the pair (U, E_f) is called an *approximation space*. In the approximation of A, the set difference $E^*(A) - E_*(A)$, defined by $E^*(A) \cap \big(E_*(A)\big)^c$ is called *boundary region*. For the example depicted in Figure 3.20 we have

$$\begin{aligned} E^*(A) - E_*(A) &= \big\{[u_1]_f, [u_2]_f, [u_3]_f, [u_4]_f\big\} \cap \big\{[u_1]_f, [u_2]_f, [u_4]_f, [u_5]_f, [u_6]_f\big\} \\ &= \big\{[u_1]_f, [u_2]_f, [u_4]_f\big\} \end{aligned}$$

A *rough set membership function* of A is then defined $\forall u \in U$ by the mapping

$$\mu_A^r(u) = \frac{\#\big([u]_f \cap A\big)}{\#\big([u]_f\big)} \tag{3.49}$$

where $\#(\cdot)$ denotes the cardinality, assuming a finite set U. We find that

$$\mu_A^r(u) = \begin{cases} 1 & \text{if} \quad u \in E_*(A) \\ 0 & \text{if} \quad u \in U - E^*(A) \\ 0 \leq \mu_A^r(u) \leq 1 & \text{if} \quad u \in E^*(A) - E_*(A)\ . \end{cases}$$

The membership function $\mu_A^r(u)$ describes the degree of possibility of u belonging to A in U. The relationship between rough set theory and possibility theory (fuzzy sets) is discussed in [18]. In particular, the upper and lower approximations $E^*(A)$, $E_*(A)$ of a fuzzy set $\widetilde{A}$ by E are fuzzy sets of U/E with membership functions defined by

$$\mu_{E^*(A)}\big([u]_f\big) = \sup\big\{\mu_{\widetilde{A}}(u)\ :\ \Gamma([u]_f) = [u]_f\big\}$$
$$\mu_{E_*(A)}\big([u]_f\big) = \inf\ \big\{\mu_{\widetilde{A}}(u)\ :\ \Gamma([u]_f) = [u]_f\big\}$$

where $\mu_{E^*(A)}([u]_f)$, $(\mu_{E_*(A)}([u]_f))$, is the degree of membership of $[u]_f$ in $E^*(\widetilde{A})$, called a *rough fuzzy set*. The accuracy of approximation of A by a rough set is calculated by

$$\mathcal{A}(A) = \frac{\#\big([u]_f \in E_*(A)\big)}{\#\big([u]_f \in E^*(A)\big)} = \frac{\#\big([u]_f \subset A\big)}{\#\big([u]_f \cap A \neq \emptyset\big)} \tag{3.50}$$

with $\#\big([u]_f \in E^*(A)\big)$ being non-empty such that $0 \leq \mathcal{A}(A) \leq 1$. The concept of linkage between two factors and a measure of accuracy like (3.50) are very important for formal modelling as only if we have achieved a synthesis with experimental data and with the elimination of information about variables (factors) that are irrelevant for the "sufficient" description of the phenomena, we achieve real understanding. If these elements are not given in a conceptual framework, the model will fail to 'explain' the phenomena and at best suggest that observed events have a reason.

Studying pairs of factors, in sections 3.1 and 3.5.2 we studied the limits of discernability of a subset of objects A belonging to the domain, or universe U. The question has been how well subset A (resp. $\widetilde{A}$) can be characterised in terms of the information available from using factors f evaluating objects in universe U. The limits of discernability of objects are due to equivalence relation E_f induced by f. The factor f plays a central role in our model realising the modelling relation (Fig. 2.1), describing the uncertainty associated with measurement and observation.

To discover cause-effect relationships among two factors f and g in F we consider the quotient set U/E_g of U w.r.t E_g. The lower approximation of the equivalence class $[u]_g \in U/E_g$ in terms of equivalence classes generated by E_f, is the set

$$E_*([u]_g) = \left\{ [u]_f \ : \ [u]_f \subseteq [u]_g \right\} . \tag{3.51}$$

Then a measure for the linkage between factors f and g (cf. (3.46), pg. 100) is given by

$$\mathcal{L}(f,g) = \frac{\#\big(\cup\{E_*([u]_g) \colon [u]_g \in U/E_g\}\big)}{\#(U)} . \tag{3.52}$$

The measure $0 \leq \mathcal{L}(f,g) \leq 1$ describes the dependency of g on f such that for $\mathcal{L}(f,g) = 0$, f and g are considered to be independent. A value close to 1 suggests a cause-effect relationship between f and g representing the conditions in which gene-expression, protein-protein or gene-gene interactions are analysed.

Remark. Note that in the present section we have considered crisp sets A, which are only a special case of a fuzzy set $\widetilde{A}$ introduced in Section 3.1. The ideas presented so far are directly applicable to nominal (categorical or qualitative, boolean or integer-valued) factors for which an equivalence relation can be established without considerations of an error or tolerance bound. If however, $X(f) = \mathbb{R}$, measurements are in a continuous space, we have the situation discussed in Section 3.5.1, leading to a non-transitive relation (3.43). Alternatively, we may quantise the $X(f)$ such that comparisons such as $f(u) = f(u')$ become again binary-valued.

3.6 SUMMARY: THEORY IN PRACTISE

Molecular biology and the biotechnology it has created, has generated vast amounts of information about properties of components and their involvement in various biological processes. However, relatively little synthesis of basic biological facts has occurred so far and the post-genome challenge is to be able to interpret and use the genome data: focus is shifting from molecular characterisation to understanding functional activity. The historical roots of analytical, reductionist paradigms are likely to be the cause for a lack of synthetic, integrative thinking required to acquire such knowledge. The research underlying the present text has been the attempt to escape the vicious circle and to contemplate about the questions of what genes do, how they interact, and whether or how integrated models of such phenomena are possible.

The fuzzy relational factor space approach developed has been based on a phenomenological theory of gene function and gene expression. That is, the model is based on the phenomenology of observations (measurements) rather than physical models of their causes on a molecular level. The interface

between an experimental context and the concept of gene function has been formally established by means of *factors*. A factor f induces an equivalence relation E_f on U and hence generates a partition in form of the quotient set U/E_f. On the basis of an one-to-one correspondence between the quotient set U/E_f and the image set or range $f(U)$ we are able to discuss the effectiveness of one factor describing the biological context and the effect of additional factors by means of equivalence classes.

Table 3.2 Summary of biological aspects in genome analysis and their formalisation using a fuzzy relational factor-space approach.

Biological Concept	**Formalisation**
Genome study, experiment	Description frame $(U, \mathcal{C}, F)$.
Sequence, gene	Object $u \in U$.
Context	Concept $C \in \mathcal{C}$.
Gene function, "phenomenon"	Extension of C in U: $\widetilde{A} \in \mathcal{F}(U)$.
Observable aspect of gene expression	Factor $f \in F$.
Representation space of f	State space $X(f)$.
Collection of factors describing C	Factor Space $\{X(f)\}_{f \in F}$.
Measurement, characterisation	State $f(u) \in X(f)$.
Gene expression, "symptom"	Repres. ext. of C in $X(f)$: $f(\widetilde{A}) \in \mathcal{F}\big(X(f)\big)$.
Known gene expression, observation	$\widetilde{B}(f) \in \mathcal{F}\big(X(f)\big)$.
Modelling relation: $\widetilde{A}$ vs. $\widetilde{B}(f)$	Feedback extension of C w.r.t. f: $f^{-1}\big(\widetilde{B}(f)\big)$.

The description frame $(U, \mathcal{C}, F)$ is suited for a variety of problems and the interpretation or semantics of the objects in U and factors in F is of utmost importance. In most cases, we shall study gene function and expression by means of some changes which, for instance, may be caused by *genome modification*. A collection of modifications to genes is called a *strain*. A strain usually leads to variations in the phenotype of the organism. Each modification to a gene in a strain is referred to as an *allele* which in turn are described as either *mutants* or *wild type*. The 'wild type' characterises the cell or organism that displays the 'typical' phenotype and/or genotype while a mutant refers to the altered or changed DNA. If a gene is considered to be a *concept* $C \in \mathcal{C}$, alleles, that is, alternative forms of a gene are then represented by $\mathcal{C}$. We may therefore use the approach in comparative studies (focussing on structural genome properties such as gene location, synteny – primarily relying on sequence information) as well as for the study of gene-expression data. These data are obtained from arrays in which each cell in an array can for example represent an ORF that is expressed in a certain context which provides clues to the function of the gene. For instance, in yeast, the context could be the study of genes effecting its growth. Transcriptome data measure an intensity of flourescent or radiating material, proportional to the level of

expression of the particular gene in question. The factors in a description frame $(U, \mathcal{C}, F)$ could therefore

- describe different conditions (contexts) in which gene expression levels are measured,
- reflect different experimental techniques in obtaining these measurements,
- represent measurements of the same context but at different levels (Transcriptome and Proteome, cf. table 1.1).

In any of these cases, the relationship or linkage between factors is of interest. While in Sections 3.1 and 3.5.2 we investigated factors w.r.t the information they represent and the effect of additional factor. In Section 3.5.4 however we introduced a measure of linkage describing causal entailment. Though the concepts and ideas, introduced in previous sections, lead to models which are sets of rules, statements about local *associations* or *dependencies* among variables, we acknowledge that the *causal problem*[28] is an ontological, not a logical question, it cannot be *reduced* to logical terms but it can be analysed with the help of formal reasoning. In the words of Bertrand Russell: "Inferences of science and common sense differ from those of deductive logic and mathematics in a very important respect, namely, when the premises are true and the reasoning correct, the conclusion is only *probable*."

So far, we have discussed 'practical problems' but only in theory. We have yet to develop means to identify, model and quantify theoretical concepts by means of sampled data. Let us summarise some formal aspects of the fuzzy relational factor-space model which are relevant to this task :

- An object is relevant to a factor if there exists a state $f(u)$, the evaluation of a factor on object $u \in U$.
- On page 53, we started off with the assumption of $V(u) = \{f \in V : R(u, f) = 1\}$ and U is *chosen* to coincide with $D(f) = \{u \in U : R(u, f) = 1\}$.
- Throughout Section 3.1 and Section 3.2, a subset $F \subset V$ was assumed to be *sufficient*, i.e., (3.4), $\forall u_1, u_2 \in U, \exists f \in F : f(u_1) \neq f(u_2)$. (See page 55).
- A *fact* was formalised by the *extension* of concept C in U, (3.5), denoted $\widetilde{A}$. The extension of C in U is a fuzzy set and $\widetilde{A}(u)$ describes the degree of relevance of u w.r.t C. (See page 56).

[28] As a definition of a 'causal law', which is not strictly bound to any specific philosophical perspective, we shall understand by a 'causal dependency' a general proposition by virtue of which it is possible to *infer* the existence of an event from the existence of another.

- We accepted that we will only be able to analyse a natural system (i.e., C or $\widetilde{A}$) by means of observable facts leading to the representation extension of C in $X(f)$, (3.6), on page 58 : $f(\widetilde{A})(x) = \bigvee_{f(u)=x} \widetilde{A}(u)$ such that $\widetilde{A} \subseteq f^{-1}(f(\widetilde{A}))$.

- As $\widetilde{A}$ is generally not known, we defined the *feedback extension*, (3.13), of C w.r.t. f based on a known representation extension $\widetilde{B}(f)$ on $X(f)$ such that the modelling relation could be described by a *rule* IF f is $\widetilde{B}$, THEN C is $\widetilde{A}$.

- The *intension* of a concept C is its description by means of a family of independent factors $G \subset F$. The feedback extension of $\widetilde{A}$ by means of independent factors in G defined, on page 64, the G-envelope, denoted $\widetilde{A}[G]$.

- For known representation extensions $\widetilde{B}(f_j)$, $f_j \in G$, we can approximate the representation extension of C on $X(f)$, $f = \bigvee f_j$, by means of the direct product, (3.22), $\widetilde{B} \approx \prod_{j=1}^{r} \widetilde{B}(f_j)$. (See page 66). This required the *cylindrical extension*, $\uparrow_{f_j}^{f} \widetilde{B}(f_j)$, (3.15), (See page 63).

- The intersection of the cylindrical extensions $\uparrow_{f_j}^{f} \widetilde{B}(f_j)$ then forms our model of the extension of C in U via (3.23) : $\widetilde{A}(u) \approx \bigwedge_{j=1}^{r} \widetilde{B}(f_j)(f_j(u))$ which may also be viewed as the rule IF f_j is $\widetilde{B}(f_j)$, THEN C is $\widetilde{A}$. (See page 66).

- Section 3.2 was concerned with approximate (rule-based) reasoning with facts represented by fuzzy sets.

- Section 3.5, reconsidered subfactors, linkage between factors. If factor g is linked to factor f at u, we can predict information about $g(u)$ via f.

- Section 3.5.4 demonstrated how probabilistic uncertainty can be integrated into the model using Bayesian belief functions.

- Section 3.5.4 also described the formal relationship of factor spaces, evidence theory with rough set theory and how cause-effect relationships can be discovered and quantified using (3.52).

Throughout the present text, the availability or use of more than one factor has been discussed in various different perspectives:

- A family of state spaces $\{X(f)\}_{f \in F}$ defines a factor space, for which the set of factors F defines a Boolean algebra. See Definition 3, page 54.

- The extension $\widetilde{A}$ of concept C in the universe of discourse U, together with the representation extension $f(\widetilde{A})$ defines a fuzzy graph $\widetilde{\mathcal{G}} = \bigvee_k f(\widetilde{A}_k) \times \widetilde{A}_k$.

- Using cylindrical extension we can get an approximate representation extension $\uparrow_g^f g(\widetilde{A})$ of C w.r.t. a more complicated factor f, by using representation extension $g(\widetilde{A})$ of C w.r.t. a simpler subfactor g. See (3.16) on page 63.
- Let $G \subset F$ be a family of independent factors, then $\widetilde{A}[G](u) = \bigwedge_j f_j(\widetilde{A})(f_j(u))$ defines the G-feedback extension of $\widetilde{A}$. The extension of $\widetilde{A}$ can then be approximated by $\widetilde{A}(u) \approx \bigwedge_{j=1}^r \widetilde{B}(f_j)(f_j(u))$. See (3.18) on page 64, and (3.23) on page 66.
- The extension of C in U can be represented as a set of if-then rules: IF f_j is $\widetilde{B}(f_j)$, THEN u is $\widetilde{A}$. The link to approximate reasoning using fuzzy systems may prove useful in implementations.
- A factor f induces an equivalence relation E_f on U and hence partitions U. Comparing the overlap between two quotient sets U/E_f and U/E_g, the concept of linkage between two factors f and g was developed in the context of two independent descriptions of the same concept. If g is linked to f, we can use f to make predictions about g. The measure of linkage $\mathcal{L}(f,g)$ can be used to identify cause-effect relations or correlations between factors. See Section 3.5.

We emphasised on the outset, that for an observer-based or phenomenological model it is of utmost importance to be precise about uncertainty. As Karl Popper demonstrated, scientific theories deal with concepts not reality. Formula and theories are so formulated as to correspond in some 'useful' way to the real world. However, this is an approximate correspondence. Mathematical forms say by *themselves* nothing about material reality. Any objective content lies entirely in the (biological, physical, ..) meaning attached *ad hoc* to the symbols appearing in mathematical formulations. There is no wrong theory, model etc. instead one may be more useful or convenient than another. The quest for precision is analogous to the quest for certainty and both precision and certainty are impossible to attain. It is therefore important to be precise about uncertainty, not to ignore it but to incorporate it in our models and theories. The presented approach addressed model uncertainty and uncertainty of the modelling process in the following ways:

Uncertainty in Modelling:

- The intension of a concept describes the properties of a concept in terms of factors. The extension of a concept is the aggregate of objects characterising it.
- Knowledge about extension $\widetilde{A}$ of concept C is gathered via measurements $f(u)$ leading to the representation extension $f(\widetilde{A})$ in $X(f)$. See Definition 6.
- Fuzziness: The relevance of an object $u \in U$ w.r.t concept $C \in \mathcal{C}$ is expressed by fuzzy set $\widetilde{A}$ in U. See (3.5) on page 56.

Model Uncertainty:

- A measure of coincidence (Def. 7), is used as a means to quantify the 'precision' to which $\widetilde{A}$ is described by $f(\widetilde{A})$.
- An approximate description of a set A in U is possible in the setting of equivalence relations and rough set theory. The quality of the approximation $\mathcal{A}(A)$ of A by a rough set is determined by (3.50), page 104. As shown by Dubois and Prade, the rough set approach can be extended to fuzzy sets $\widetilde{A}$.
- Partial evidence about A in U can be integrated by means of a probability measure $m\colon \mathcal{P}(U) \to [0,1]$ leading to evidential reasoning. See Section 3.5.4.

Uncertainty in Data and Measurement:

- We allowed for nominal (categorical or qualitative) data as well as real numbers.
- Imprecise (interval valued) data can be considered. The membership function $\widetilde{B}$ in $X(f)$ is estimated as the one-point coverage function (cf. [78, 37]).
- For sampled data, assumed to follow a probability law, a bijective transformation between histograms and possibility distributions can be used to determine fuzzy set $\widetilde{B}$ (cf. [77]).
- In Section 3.5.1 measurement errors were introduced. For an error or tolerance bound ε two evaluations of a factor on two objects u and u' are indistinguishable and induce a relation $R_\varepsilon = \{(u,u')\colon |f(u)-f(u')| \leq \varepsilon\}$ which in not transitive. Non-transitivity motivates fuzzy relations.

We have shown that a comprehensive (mathematical) theory for genome analysis can be realised (on paper...) using fuzzy mathematics and system theory. We considered two particular aspects: the study of genome structure, location and dependency of genes based primarily on sequence information and the study of gene function using gene-expression data. Our approach is characterised by

▷ *Scalability*: The concept of factors spans a range of magnitudes (see Figure 1.1).

▷ *Flexibility*: The theory is applicable to a wide range of problems.

▷ The possibility to discuss the modelling process itself; hence allowing us to be precise about uncertainty.

For a software implementation we established formal relationships between a factor-space model and

- ▷ Fuzzy Rule-Based Reasoning,
- ▷ Rough Set Theory,
- ▷ Evidence Theory.

Each approach to knowledge representation for itself as hybrid paradigms have lead to a wide range of successful applications and hence promise a feasible way forward to verify the presented fuzzy relational theory with real data from biological databases.

The formal mathematical model was developed in the context of the modelling relation (Figure 2.1, page 23) describing the process by which we model natural systems and reason about natural laws. The modelling relation however is more than merely a representation of modelling, it represent the scientific approach in general and the process by which individuals learn in particular. In Section 2, the concept of 'differentiation' was described as central to science, and modelling. Previous discussions Section 3.5, considered equivalence relations arising from observation only, i.e., different modes of observation by means of more than one factor to discern objects. An equivalent approach is to observe a system with the same factor but under changed conditions. Removing parts of system, perturbation or excitation produce a change which allows us to study a system's behaviour. The discrepancy between behaviours helps us to determine the function of a systems or its components. The discrepancy of components helps us to determine the system's structure. The experiments involved may be simply comparative, in space or in time.

The dualism of understanding by observation (discerning) and manipulation through changes is an interesting aspect for the history of science. Biology itself currently undergoes a transition away from a field of observation and contemplation. Its researchers have been primarily fascinated and motivated by the complexity of behaviour and structure while recently biological research is argued for with the possibility of manipulation. The excitement, marvel and respect, arising from contemplating about nature is well summarised by Arthur Schopenhauer's observation that 'anyone can squash a bug but all professors of this world couldn't build one'.

In the present chapter, we tried to develop a conceptual framework of genomics. The two central questions of genomics, regarding the genes' biological functions, i.e., relationships between groups of genes, and the analysis of interactions between genes and proteins, both have their equivalent in mathematical equivalence relations and linkages between observable factors. The aim of a conceptual framework of genomics, composed of these two mathematical concepts, is to help explain unknown relationships, predict or simulate and to help design experiments, telling us which variables to measure and why. As the mathematician David Hilbert once said, 'there is nothing more prac-

tical than a good theory'. And yet we must be aware of Albert Einstein's caution that 'as far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality.' Although the mathematical structures we employ to encode natural systems, are not in themselves the reality of the natural world, I believe that they are the only key we posses to that reality.

4

Systems Biology

In this chapter we discuss whether systems biology is the reincarnation of systems theory applied in biology. There is necessarily some repetition with parts of previous chapters in order to lead us to Rosen's Metabolism-Repair (M,R)-systems which form the main focus of this chapter[1].

With the availability of quantitative data on the transcriptome and proteome level, there is an increasing interest in formal mathematical models of gene expression and regulation. International conferences, research institutes and research groups concerned with *systems biology* have appeared in recent years and systems theory, the study of organization and behavior *per se*, is indeed a *natural* conceptual framework for such a task. This is, however, not the first time that systems theory has been applied in modelling cellular processes. Notably in the 1960s systems theory and biology enjoyed considerable interest among eminent scientists, mathematicians and engineers. Why did these early attempts vanish from research agendas? Here we shall review the domain of systems theory, its application to biology and the lessons that can be learned from the work of Robert Rosen. Rosen emerged from the early developments in the 1960s as a main critic but also developed a new alternative perspective to living systems, a concept which deserves a fresh look in the post-genome era of bioinformatics.

[1]Parts of this chapter have been published in [80].

4.1 OVERVIEW

We see an ever-increasing move towards inter and trans-disciplinary attacks upon problems in the life-sciences. The reason is the diversity of organization and behavior in natural systems. The size of data sets and complexity of patterns hidden in them has led to a renewed interest in mathematical techniques that allow us to identify formal models of natural systems. The next step in the post-genome era is not simply assigning biological function to identified genes but to analyze the organization and control of genetic pathways. These pathways are of course dynamic systems; non-linear, adaptive and anticipatory systems to be precise.

Systems biology is an emerging field of biological research that aims at a system-level understanding of genetic or metabolic pathways by investigating *interrelationships* (organization or structure) and *interactions* (dynamics or behavior) of genes, proteins and metabolites. Recently, international conferences, institutes [25, 30], research groups and articles [30], focussing on systems biology, have appeared. The reason for this renewed interest in systems thinking is the rapid technological advance in the area of genomics. Genomics is the field of biological research taking 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 and, ultimatively, the organism. Crossing several scale-layers from molecules to organisms, we find that organisms, cells, genes and proteins are defined as complex structures of *interdependent* and subordinate *components* whose relationships and properties are largely determined by their function in the whole. This definition coincides with the most general definition of a system as a set of components or objects and relations among them [33]. Systems theory is then the study of organization and behavior *per se* and a natural conclusion is therefore to consider systems biology as the application of systems theory to genomics.

The idea to use systems theory in biology is however not new, notably in the 1960's a number of eminent researchers took a systems approach to 'search for general biological laws governing the behavior and evolution of living matter in a way analogous to the relation of the physical laws and non-living matter' [1, 41, 4]. It was the transfer of ideas from physics to biology and the perception that biological systems are a special case of physical systems that led to criticism which cumulated in the most comprehensive discussion of the limitations of 'classical' systems biology in the work of Robert Rosen [59, 60, 61, 62]. In the following sections, we review the need for mathematical modelling, the usual approaches to modelling biological systems and problems arising from them. In this paper we will focus on Rosen's relational biology, 'metabolic-repair' (M,R)-systems, his discussion of anticipatory behavior and causality. We show that, for metabolism and repair defined as mappings, replication is implicitly defined. Anticipatory behavior or intrinsic control is

realized through the boundary conditions of the repair and replication map. Finally, it can be shown that the category which defines the (M,R)-system is rich enough in entailment to allow the repair and replication maps to be entailed by something and hence avoiding a finality argument when discussing causal entailment.

4.2 THE CASE FOR MATHEMATICAL MODELLING

The engineering sciences are a good example of how mathematics has been used effectively in a wide range of applications. One could argue that many biologists find themselves now in a similar situation to engineers about six decades ago when they were faced with the need to analyze and control complex dynamic systems for which empirical means are inappropriate. Also, both species, engineers and biologists are not born as mathematicians. Engineers have learned to use mathematics towards their ends and a symbiosis of researchers from both areas should allow both to advance successfully. For the engineer, the underlying strategy is to represent the natural system by a set of random- and/or state-variables and then to investigate relationships among those variables within a formal system (Figure 2.1). This approach cumulates into a philosophy whereby, as Henri Poincaré suggested, "the aim of science is not things in themselves but the relations between things; outside these relations there is no reality knowable."[51, page xxiv]

The importance of what we now call systems biology was pointed out by Norbert Wiener in his book *Cybernetics, on Control and Communication in the Animal and the Machine*, published in 1948 [76]. In 1970, cybernetics or feedback regulatory mechanism on a molecular level were described by Jacob and Monod [28, 45] who investigated regulatory proteins and the interactions of allosteric enzymes in particular. Organisms as a whole are self-regulating, adaptive and anticipatory systems and numerous examples have been published. While the control of physiological mechanisms requires the processing of *information*, the actual processes are sustained by *energy* obtained from the environment. The acquisition, transfer and utilization of energy has subsequently been seen as a major component in the analysis of biological systems [75]. Systems biology has a past and the the books by Ashby [1] and Bertalanffy [4] are a 'must read' for anyone attracted to the area of systems biology. Bertalanffy provides a general introduction of system theory but also reviews applications in biology with a discussion on models of open systems and organisms considered as physical systems. For an up-to-date account of the systems sciences, including a historical perspective, the reader is referred to Klir's book [33] and the Principia Cybernetica Web [52]. Specifically referring to applications in biology, the volume *Systems Theory and Biology* edited by Mihajlo Mesarović [40] is valuable. Mihajlo Mesarović initiated and developed one of the most comprehensive mathematical systems theories [42, 43]. The

most extensive discussion of systems thinking in biology is James G. Miller's book on a 'general theory of living systems' [44]. Miller provides the most detailed account of living systems in eight levels of increasing complexity - from molecules to cells, organs, organisms and societies. Reality is described as a continuous dynamic process, best represented as a system of systems and natural systems are studied as a structure of processes evolving through spatio-temporal events. The conclusion is that despite the endless complexity of life, it can be organized and repeated patterns appear at different levels. Indeed, the fact that the incomprehensible presents itself as comprehensible has been a necessary condition for the sanity and salary of scientists.

4.3 CAUSING PROBLEMS

The principal purpose of mathematical models applied in the natural sciences is to identify sets of rules, statements about local *associations* or *dependencies* among variables. In genomics, mathematical models may be expected to not only *describe* associations but also to *explain* dependencies among genes. A 'causal law', which is not strictly bound to any specific philosophical perspective, is then understood as a 'causal dependency' a general proposition by virtue of which it is possible to infer the existence of an event from the existence of another. It is the explanatory aspect of mathematical modelling which leads us to the limits of systems biology but it is also the most exciting aspect of the developments in the post-genome area. We find that the 'causal problem' is an ontological, not a logical question, it cannot be reduced to logical terms but it can be analyzed with the help of formal reasoning. In the words of Bertrand Russell: "Inferences of science and common sense differ from those of deductive logic and mathematics in a very important respect, namely, when the premises are true and the reasoning correct, the conclusion is only *probable*." [63, page 353]

The first comprehensive theory of causation was Aristotle's. It distinguishes four types of cause: the material cause (or stuff), the formal (formative) cause (or shape), the efficient cause (or force) and the final cause (or goal). For a formal logical system, given an 'effect', say proposition P, axioms correspond to the material cause of P, production rules are understood as the efficient cause of P and the specification of particular sequences of production rules or an algorithm is identified as the formal cause. For a dynamic system a state can itself be entailed only by a preceding state. If for a chronicle $\{(n, f(n))\}$ we ask *why* the n^{th} entry gives the particular value $f(n)$, the answer is *because* of the initial condition $f(0)$, i.e., $f(0)$ is the material cause; and *because* of a state transition mapping T for which $f(n+1) = T(f(n))$, i.e., T corresponds to the efficient cause; and *because* of exponent n from which $f(n)$ is obtained by iterating the transition map n times beginning with $f(0)$; i.e., n refers to the formal cause. As shall be discussed in further detail below, in Rosen's

relational biology, for a component $f\colon A \to B$, such that $a \mapsto f(a)$, the question "why $f(a)$?" is answered by "because f" and "because a". In other words, "a entails $f(a)$" or formally $f \Rightarrow (a \Rightarrow f(a))$. Here f corresponds to the efficient cause of ("effect $f(a)$"), and a refers to the material cause of $f(a)$. One of Rosen's achievements is that he introduced a formalism rich enough in entailment, to allow final causation without implying teleology. The conceptual framework in which he developed his *relational biology* is category theory [58, 36].

4.4 TOWARDS A RELATIONAL BIOLOGY

The problems of applying systems theory in biology can be summarized by a) the difficulty of building precise and yet general models, b) the 'openness' of biological systems, the fact that these systems are hierarchical and highly interconnected, and c) that models based on differential equations cannot represent anticipatory behavior as present in cellular processes.

Modelling systems with sets of first-order differential equations,

$$\frac{\mathrm{d}f_j}{\mathrm{d}t} = \phi_j(f_1, \ldots, f_r) \ , \qquad j = 1, \ldots, r$$

the rate of change of observable (state-variable) f_j depends *only* on present and past states but cannot be dependent upon future states. In other words, these systems can only be *reactive* but not *anticipatory* [60]. The reactive paradigm embodies one of the most important assumptions of science, effects should not precede their causes. And yet simple biological systems suggest the notion of *self-reference*, an implicit model of knowledge of itself. The following example of a biosynthetic reaction network is due to Robert Rosen [60] (See also [12]). Let metabolites A_i represent the substrates for the enzyme E_i that catalyzes it at stage i. As illustrated in Figure 4.1, the initial substrate A_0 activates the enzyme E_n (i.e., increases its reaction rate). Under the foregoing hypotheses, with concentration A_0 at time t the concentration of A_n at some future time is predicted in order to maintain homeostasis in the pathway. The ambient concentration of A_0 serves as a *predictor*, which in effect 'tells' the enzyme E_n that there will be an increase in the concentration A_{n-1} of its substrate, and thereby pre-adapts the pathway so that it can deal with the expected changes.

The second problem faced by representing cellular processes with sets of linear differential equations is captured by Zadeh's uncertainty principle [34]:

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

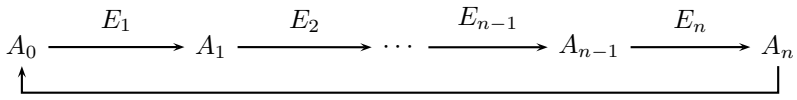


Fig. 4.1 An anticipatory chemical reaction network.

The problem is that perturbations to cells have multi-gene / multi-transcript / multi-protein responses, 'closing' the system, i.e., restricting the model to a small set of variables inevitably leads to an often unacceptable level of uncertainty in the inference.

The tradition of describing cellular systems in terms of energy and masses with forces acting on them is rooted in the realm of Newtonian mechanics. In this context a system is closed by internalizing external influences through added state variables and more parameters to the system. Take for example the simplest of dynamical system, a single particle moving along a line under the action of a constant force, the motion is governed by Newton's Second Law, which defines the force F acting on a mass point m to be the rate of change of momentum $(m \cdot v)$:

$$F = m \cdot \frac{\mathrm{d}v}{\mathrm{d}t} \;=\; m \cdot \frac{\mathrm{d}^2 x}{\mathrm{d}t^2} \;,$$

with v denoting the velocity which, in turn, is defined as rate of change of position or displacement x from some origin of coordinates. 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:

$$F(x, v) = -\theta \cdot x \;,$$

where θ is a parameter specific to the system under consideration. Rewritten as a set of first-order differential equations, this system has two *state-variables*, f_1 denoted x and f_2 denoted by v, such that

$$\frac{\mathrm{d}x}{\mathrm{d}t} = v \quad \text{and} \quad \frac{\mathrm{d}v}{\mathrm{d}t} = -\frac{\theta}{m} \cdot x \;.$$

The model is deterministic in that the object's state at time t is fully determined from the initial conditions (if known) and therefore permitting the prediction of future states by integrating the set of differential equations. Newton's laws of motion, which state that the acceleration of an object is directly proportional to the force acting on it and inversely proportional to its mass, imply that the future behavior of a system of bodies is determined completely and precisely for all time in terms of the initial positions and velocities of all

the bodies at a given instant of time, and of the forces acting on the bodies. These forces may be *external forces*, which arise outside the system investigated, or they may be *internal forces* of interactions between the various bodies that make up the system in question. Rosen described the response of a system to forces as the 'inertial' aspect while the exertion of forces by the system corresponds to the system's 'gravitational' aspect. He suggested a shift attention from exclusively 'inertial', i.e., structural aspects such as the DNA molecule and its sequence, to 'gravitational' concepts. Instead of concerning us with material causation of behavior, manifested in state sets, he suggested formal and efficient causations as the focus of attention. Such a shift of perspective is exemplified in category theory, Rosen's preferred language to discuss these problems in the abstract, by studying mappings between sets (of objects) rather than analyzing the objects themselves.

Phenotypes are what we can observe directly about organisms. They are tangible, material properties that we can measure, can compare and experiment with. The phenotype is seen as being 'caused' or 'forced' by the genotype. As Rosen points out in [62], the phenotype–genotype dualism is allied to the Newtonian dualism between states and the forces that change the states. In Aristotelian language, the states represent material causation of behavior, while the forces are an amalgam of formal and efficient causation. Biological phenotypes, considered as material systems, are open. They are open to 'forcing' by genes as well as open to interactions with their environment. To study an open system it is therefore necessary to consider the 'outside', the environment in order to understand what is going on 'inside'. The Newtonian paradigm, underlying the traditional approach to modelling biological systems, is frequently seen as synonymous with reductionism and its failure to supply the whole from its parts. On the basis of this analysis and continuing the work of Rashevsky, Rosen argued his case for a new approach, called *relational biology*. He emphasized that we must look for principles that govern the way in which physical phenomena are organized, principles that govern the *organization* of phenomena, rather than the phenomena themselves. Relational biology is therefore about organization and describes entailment without states. The association of energy or matter, described by states and dynamical laws, is to be replaced by the description of a system in terms of its components, their function and contribution to the organization of the system. An example of this approach for molecular systems is Rosen's concept of Metabolism-Repair or (M,R)-systems.

4.5 METABOLISM-REPAIR SYSTEMS

Driven by technological advances and the sequencing of genomes, at present, more hypotheses are generated than tested. However, with the availability of data, biologists will soon return to refined biological questions, "zooming

in" to specific genetic pathways. With the boom in bioinformatics, the attempts to explain genetic systems are likely to proceed from the Cartesian metaphor, viewing organisms as performing computations, describing biological principles in the same way as *machines* are. This tradition has its roots in Newtonian mechanics and formal logic, embodied in reductionism. As we witness a shift of focus from molecular characterization to an understanding of functional activity in genomics, this strategy is prone to repeat historical failures as outlined in Rosen's *'Comprehensive inquiry into the nature, origin, and fabrication of life'* [61]. As bioinformaticians dream of *in silico* models of cellular systems, Rosen developed a new biology *on paper*. Starting from the *modelling relation*, illustrated in Figure 2.1, he began by considering two natural systems N_1 and N_2 as analogues when they realize a common formalism F. This relation of analogy between natural systems is then independent of their material constitution. The formal system F is *relational*, consisting of a set of formal, interrelated, *components*. Any two natural systems that realize this formalism are said to manifest a common *organization*. In relational biology a component is defined by a *mapping*

$$f : A \rightarrow B$$

where the 'identity' of the component is embodied in the mapping itself, while the influence of surrounding components of the natural system N and the external environment are embedded in the specific *arguments* in the domain A on which the mapping can operate.

Section 4.4 introduced the anticipatory character of biological systems. The basis for anticipatory behavior is a form of self-reference or internal modelling. A *cell* is a good example of a self-referential system. We can describe a cell functionally as consisting of two major functional components, reflecting the morphological partition between nucleus (genome) and cytoplasm (phenome). The *metabolic* or ergonic component represents its basic chemical activity through the acquisition, transfer and utilization of energy. The *repair* or cybernetic component ensures continued viability of the cell in the face of external disturbances. The latter requires the processing and utilization of information to permit the control of what the cell does and characterizing its temporal characteristics. Essential for the maintenance of life, both components are closely interrelated in jointly sustaining the steady state [75]. Rosen devised a class of relational cell models called Metabolism-Repair (M,R)-systems to characterize the minimal organization a material system would have to manifest or realize what is called a *cell* [61]. The present section addresses Rosen's answer to the problems of causation, discussed in Section 4.3, and anticipatory behavior, described in Section 4.4. We are going to review Rosen's arguments and show (in the abstract) that the presence of 'metabolism' and 'repair' components imply the existence of a 'replication' principle. The key point is that replication comes without infinite regress in modelling and hence allows the discussion of final causation while avoiding the

explanation of phenomena by the purpose they serve rather than by postulated causes (teleology). To achieve this, we require a conceptual framework rich enough in entailment – such as category theory.

Let A represent the set of environmental inputs to the cell, while B is the set of outputs, i.e, products the cell is capable of producing. The mapping f could be described as an abstract 'enzyme', which converts substrate $a \in A$ into 'product' $b \in B$:

$$\begin{aligned} f : \quad & A \to B, \qquad\qquad f \in \mathcal{H}(A,B) \\ & a \mapsto f(a) = b \ . \end{aligned} \tag{4.1}$$

Further, let $\mathcal{H}(A,B)$ be the *set* of metabolisms which are realizable by the cell, i.e., a set of mappings from A to B. As pointed out by Casti [11], the set of physically realizable cellular metabolisms $\mathcal{H}(A,B)$ is determined by various physicochemical constraints and the classical Newtonian machinery has been used to capture many aspects of the cell's metabolic activity in respect of the mapping f above. However, both Rosen and subsequently Casti have argued that these formalisms lack a structure to account for *repair* and *replication*. The purpose of repair is to stabilize cellular metabolic activity against fluctuations and disturbances in both its environmental inputs and in the metabolic map f itself. In other words the repair is to copy f while we refer to replication as the process of copying the repair mechanism.

To arrive at a repair mechanism we consider the following diagram:

$$A \xrightarrow{\ f\ } B \xrightarrow{\ g\ } C \ . \tag{4.2}$$

In the diagram, a entails $f(a)$ and referring to the discussion in Section 4.3 we can answer the question "why $f(a)$?" in two ways: because a entails $f(a)$ and because f acting on a entails $f(a)$. We can summarize the entailment in the diagram by

$$\forall a \in A, \quad f \Rightarrow \big(a \Rightarrow f(a)\big) \quad \text{and} \quad g \Rightarrow \big(b \Rightarrow g(b)\big) \quad \forall b \in B \ .$$

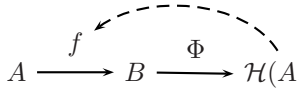
If an element $b \in B$ is entailed, then it must lie in the range of mapping f and we can write $f(a) = b$ for some element a in the domain of f and obtain

$$g \Rightarrow \Big(f(a) \Rightarrow g\big(f(a)\big)\Big) \ .$$

Suppose the set C in the diagram denotes the collection of mappings from A to B, $\mathcal{H}(A,B)$, we then find that g in fact generates a new f for any $b \in B$. In other words, $g(b)$ is itself a mapping such that g entails f:

$$g\big(f(a)\big) = f \ .$$

In this case we denote this *'repair map'* by Φ and illustrate the repair process by the following augmented diagram:



To allow some form of internal control, the repair map Φ converts the abstract products b into new versions of f:

$$\Phi : \quad B \to \mathcal{H}(A,B) \ . \tag{4.3}$$

For any specific activity, we denote the metabolism for which the cellular process is designed by f^*; i.e., in the absence of disturbances, given the environmental input $a^* \in A$, f^* produces the cellular output $b^* \in B$. If there is a disturbance to the metabolic function f^* or a change from the environment a^*, the cell 'repairs' the situation by generating new f^* for any b^*. The repair or control is implicit in the *boundary condition* of the repair map Φ: If there is neither a change from the metabolic map f^* nor from the environment a^*, then Φ ought to produce f^*:

$$\Phi_{f^*}\left(b^*\right) = f^* \ ,$$

stabilizing the cell's metabolic behavior in response to external influences and/or errors. While in the simple diagram (4.1), representing a *metabolism*, we could answer the question of "why $f(a)$", f itself was unentailed. The finality argument would be to answer "because f", f is to bring A into B and yet f is itself unentailed if we had not Φ in place. However, with the introduction of of the *repair* function $\Phi \Rightarrow (f(a) \Rightarrow f)$, the question "why f?" is answered "because Φ", Φ being the efficient cause of f and "because $f(a)$", where f is entailed by its value, the material cause.

The construction of the repair map immediately poses the question to what replicates Φ? One solution is to add yet another function to the diagram (4.2) but this would lead to an infinite regress in the discussion of causal entailment. The cell's metabolic processing apparatus, through information stored in the DNA, allows replication and it was Rosen's major achievement to show that, using category theory [58, 36], replication is in fact already built into the scheme outlined in diagram (4.2). Although we can add a replication map to the diagram, we do not need to argue for this map through an addition to (4.2) as it already implicitly exists.

To arrive at this conclusion, we view the quadruple (A, B, f, Φ) as a simple (M,R)-system on the *category* $\mathbf{C}$. A category comprises a collection of *objects* such as A, B and associated *arrows* (mappings) such as for example $f\colon A \to B$, where A is the *domain* of f and B its *co-domain*. The collection of all mappings with domain A and co-domain B is denoted $\mathcal{H}(A, B)$. We suppose

that **C** is a concrete category, i.e, its objects are structured sets and its arrows are mappings compatible with their structure. If **C** is closed under cartesian products, i.e., if A, B, C, D are objects of **C**, $f \in \mathcal{H}(A,B)$, and $g \in \mathcal{H}(C,D)$ are maps of **C**, then $A \times C$ and $B \times D$ are objects of **C**, and $f \times g \in \mathcal{H}(A \times C, B \times D)$, where $(f \times g)(a,c) = \big(f(a), g(c)\big)$, then for $\mathcal{H}(A,B)$ defining an object in **C**, we introduce for this special case a new notation:

$$B^A = \{f \mid f\colon A \to B\} \ .$$

In (4.3) above we have in fact assumed that $\mathcal{H}(A,B) = B^A$ is an object in the category to which A and B belong because only then B^A can be the range of another mapping in the category and hence can be entailed within the category. In other words, for Φ to entail f, the *exponential* (function set or map object) B^A must exist. Note that B^A does not necessarily exist as an object in **C**, there is, for example, no analogous construction in the category of monoids. If B^A exists as an object of the category **C**, it is associated with the existence of a special *evaluation mapping* e:

$$\begin{aligned} e_f \ : \quad \big(B^A \times A\big) \ &\to \ B, \\ (f,a) \ &\mapsto \ e_f(f,a) = f(a) \ . \end{aligned} \tag{4.4}$$

Note that we use subscript f in e_f not to denote a dependency on f but to distinguish it from evaluations associated with maps other than f.

Returning to our cellular (M,R)-system

$$A \xrightarrow{\ f\ } B \xrightarrow{\ \Phi\ } B^A \ , \tag{4.5}$$

the map Φ is an element of the set of mappings from B to the set of mappings from A to B:

$$\Phi \in \big(B^A\big)^B \ .$$

For this *set* to exist as an *object* in the category **C**, following the general model (4.4), there then must exist the evaluation map

$$\begin{aligned} e_\Phi \ : \quad \Big(\big(B^A\big)^B \times B\Big) \ &\to \ B^A, \\ (\Phi, b) \ &\mapsto \ e_\Phi(\Phi, b) = \Phi(b) \ . \end{aligned} \tag{4.6}$$

The existence of the evaluation map e_Φ can be explained as follows. The value $f = \Phi(b)$ can be viewed as depending on *two* things: b as well as Φ. Although we don't usually think that way with functions, there is no reason why Φ is fixed in the setting of (M,R)-systems. We can express this dependency on both Φ and b as a two-valued mapping

$$(\Phi, b) \mapsto f \ .$$

Suppose we want to *evaluate* this map, which we denote by h for now, we can describe this as a two-step process which effectively turns the mapping $h(\Phi, b)$ of two variables into a map $H(\Phi)$ of one variable Φ but with values $H(\Phi)$ which are a function of the second variable b. The formal definition of this map H reads

$$\big(H(\Phi)\big)(b) = h(\Phi, b), \quad \text{where} \quad \Phi \in \big(B^A\big)^B, \text{ and } b \in B\ . \tag{4.7}$$

Here each value $H(\Phi)$ is a function of b, hence an element of the exponential set

$$\big(B^A\big)^B = \big\{\Phi \mid \Phi\colon B \to B^A\big\} \quad \text{such that} \quad H\colon B \to \big(B^A\big)^B\ .$$

In formula (4.7) on the left-hand side the mapping $H(\Phi)$ is evaluated at argument b and h may therefore be called an *evaluation map* and denoted by e_Φ, leading us to the definition in (4.6).

Our reasoning so far can be summarized as follows. For Φ, the repair of f, being entailed by something (being replicated), it is required that the set of mappings from B to B^A exists as an object in $\mathbf{C}$. Then, if such a map object (exponent) exists, it is associated with the evaluation map e_Φ. The evaluation map in turn was explained by the bijection

$$\frac{h\colon \big(B^A\big)^B \times B \to B^A}{H\colon B \to \big(B^A\big)^B}$$

between functions h in two variables and those H in one variable but which maps into $\big(B^A\big)^B$, the space in which Φ resides! In other words, given the metabolic function $f\colon A \to B$, and repair map $\Phi\colon B \to B^A$, these imply the replication of Φ. With replication of Φ in place, we can introduce a *replication map*, denoted Υ,

$$\Upsilon\ :\quad B^A \to \big(B^A\big)^B\ , \tag{4.8}$$

such that Φ is entailed by f. As previously defined for the repair map, the boundary condition for a stable operation is $\Upsilon(f) = \Phi_f$. The boundary conditions are important as they define the (M,R)-systems as a controlled process. In conventional control engineering the existence of a separate control component is assumed. The control action is an external influence on the process and we may refer to this type of control as extrinsic (exogenous). For (M,R)-systems there is no direct control input and the separation between controller and process is not recognizable (intrinsic or endogenous control). Instead the 'anticipatory regulation' is implicit in the boundary conditions for Φ and Υ. The boundary conditions imply an internal self-model of the cell. Given A, B and $\mathcal{H}(A, B)$, it is possible to directly construct the maps Φ_{f^*} and Υ_{f^*}, i.e., repair (of metabolism f) and replication (of the repair map Φ) emerge 'naturally' from the existence of an abstract metabolic component. An argument in support of theoretical or mathematical biology is that such

results, abstract they may be, are neither the outcome of *in vivo*, *in vitro* or *in silico* analysis but can also be obtained, *on papyrus*... .

We can realize a (M,R)-system in different ways and initially automata theory was considered. However as demonstrated by John Casti [11], since Rosen introduced the concept, considerable advances in the mathematical theory of dynamic systems should enable us to take his ideas further. In [10], Casti developed a theory of *linear* (M,R)-systems. In the model above we can consider a as an input time-series leading to output b. The input/output space A and B are then finite-dimensional vector spaces whose elements are sequences of vectors from $\mathbb{R}^m$ and $\mathbb{R}^p$ respectively:

$$\begin{aligned} A &= \{a\colon a = [u_0, u_1, \ldots, u_N]\}\,, \quad u_i \in \mathbb{R}^m, \\ B &= \{b\colon b = [y_1, y_2, y_3, \ldots]\}\,, \qquad y_i \in \mathbb{R}^p\ . \end{aligned}$$

Mathematical causation is acknowledged by the fact that the first output appears one discrete time step after the first input. If f is further assumed to be linear and constant (autonomous), we can express the relationship between cellular inputs and outputs by the following equation:

$$y_t = \sum_{i=0}^{t-1} A_{t-i} u_i, \quad t = 1, 2, \ldots, \tag{4.9}$$

where $A_k \in \mathbb{R}^{p\times m}$ denotes the coefficient matrix which characterizes the process.

The (M,R)-systems consists of $f\colon A \to B$ such that $f(a^*) = b^*$ plus $\Phi\colon B \to B^A$ such that $\Phi(b^*) = f^*$. With a linear realization (4.9) we are now in a position to investigate how the (M,R)-system restores or stabilizes disturbances in the cellular environment a and/or metabolic map f. A change in the external environment, $a^* \to a$, for a fixed metabolic map f^* leads to $f^*(a) = b^{new}$ subsequently to $\Phi(f^*(a)) = f^{new}$. For a stable process, we require that $f^*(a) = f^*(a^*)$ or $f^{new}(a) = b^*$ for the cell to recover fully from the disturbance. The cell's metabolic activity would be permanently changed to f^{new} if $\Phi(f^{new}(a)) = f^{new}$. If we had $\Phi(f^{new}(a)) = f^*$, then the cell's metabolism would only undergo periodic changes cycling back and forth between f^* and f^{new}.

In case of a fixed environment a^*, fixed repair map Φ_{f^*} with a disturbance $f^* \to f$, we require $\Phi_{f^*}(f(a^*)) = f^*$ in order to restore the original design-metabolism f^*. This in fact describes a map $f \mapsto \Phi_{f^*}(f(a^*))$. Let us denote this map as follows

$$\begin{aligned} \Psi_{f^*,a^*}\ :\quad B^A &\to B^A, \\ f &\mapsto \Phi_{f^*}\big(f(a^*)\big)\ . \end{aligned}$$

We may find that for some disturbances f the repair mechanism stabilizes the system to $\Phi_{f^*}(f(a^*)) = f^*$ but in some cases the system could settle for the new metabolism f such that $\Phi_{f^*}(f(a^*)) = f$. This situation is represented by fixed points of the map Ψ_{f^*,a^*}. One such fixed point is of course f^*, for which our basic system is working normally such that $\Phi_{f^*}(f^*(a^*)) = f^*$.

In [11, 12], Casti addresses other biological questions such as mutations and Lamarckian inheritance. We may conclude that Rosen's somewhat abstract formulation of (M,R)-systems, initially argued for by calling upon category theory and thereby allowing us to reason about more fundamental properties of cellular systems, has also more 'applied' formulations in form of sequential machines and linear dynamic systems. The formal tools required for such an analysis are familiar to control engineers. John Casti described various properties of such systems and established further links of these ideas to a number of other areas of science and engineering [10, 12]. The, for many, unexpected link between biological questions and engineering analysis should encourage control engineers in particular to take an interest in systems biology. We can expect that over the coming years new technology will allow us to measure gene expression in time. Similar approaches to those discussed here should then be developed to study gene interactions.

4.6 CONCLUSIONS AND DISCUSSION

The principal aim of systems biology is to provide both, a conceptual basis and working methodologies for the scientific explanation of biological phenomena. System theory is not a collection of facts but a way of thinking, which can help biologists to decide which variables to measure and to validate their 'mental models'. Frequently it is the process of formal modelling rather than the mathematical model obtained, which is the valuable outcome. In engineering it is a common experience that we often learn most from those models that fail. The purpose of a conceptual framework, is therefore to help explain unknown relationships, to make predictions and to help design experiments, suggesting to us which variables to measure and why. Or, as the mathematician David Hilbert once noted, we might think that 'there is nothing more practical than a good theory'.

The need for mathematical models becomes apparent as we begin to analyze the organization and control of genetic pathways. The complexity of molecular processes combined with the difficulties in observing them and measuring quantitative data, leads inevitably to uncertainty in their analysis. Mathematical models, providing sufficiently accurate numerical predictions, are possible in some cases as demonstrated in the areas of metabolic engineering and control. With applications in biotechnology the inner structure of models in this area is less important than the ability to replicate observable phenom-

ena in simulations. If however, on the other hand, we are trying to answer more fundamental questions regarding the mechanisms, principles or causal entailment in genetic pathways, we find that the ancient problem of causality haunts us once again.

Differential equations may be used to model a specific form of causal entailment in natural systems, the equations by themselves however do not state that changes are *produced* by anything, but only that they are either *accompanied* or *followed* by certain other changes. Considering $\mathrm{d}f/\mathrm{d}t = \phi(t)$ or equivalently $\mathrm{d}f = \phi(t) \cdot \mathrm{d}t$, it merely asserts that the change $\mathrm{d}f$ undergone during the time interval $\mathrm{d}t$ equals $\phi(t) \cdot \mathrm{d}t$. The notion of causality is not a syntactic problem but a semantic one; it has to do with the interpretation rather than with the formulation of theories or formal systems. In other words, hypothesizing causal entailment in general, and gene/protein interactions in particular, remains a task of the biologist, possibly supported by his *choice* of mathematical model (conceptual framework). As problems of genomics become conceptual as well as empirical, and models are expected to explain principles rather than just simulating them, we are therefore likely to witness interesting debates on the merits of alternative theories.

Scientific theories deal with concepts, not with reality and mathematical models are representations, not reflecting what things are in themselves. All theoretical results are derived from certain formal assumptions in a deductive manner. In the biological sciences, as in the physical sciences, the theories are formulated as to correspond in some useful sense to the real world, whatever that may mean. Energy or matter is the primary object of physics. Its study in the phenomenal world is based on changes and for anything to be different from anything else, either space or time has to be pre-supposed, or both. Immanuel Kant identified the concepts of space, time and causality as *a priori* and therefore conditional for experience. Changes in space and time are the essence of causal entailment and as the philosopher Arthur Schopenhauer discovered, the subjective correlative of matter or causality, for the two are one and the same, is the *understanding*. "To know causality is the sole function of the understanding and its only power. Conversely, all causality, hence all matter, and consequently the whole of reality, is only for the understanding, through the understanding, in the understanding" [38]. In his famous essay "What is life?" [66], the physicist Erwin Schrödinger, comes to the conclusion that "our sense perceptions constitute our sole knowledge about things. This objective world remains a hypothesis, however natural", echoing Albert Einstein observation that 'as far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality.'

The work of Robert Rosen is important in that he not only identified the weaknesses of our common approach to represent natural systems but he also outlined a possible way to transcend the reactive paradigm in order to

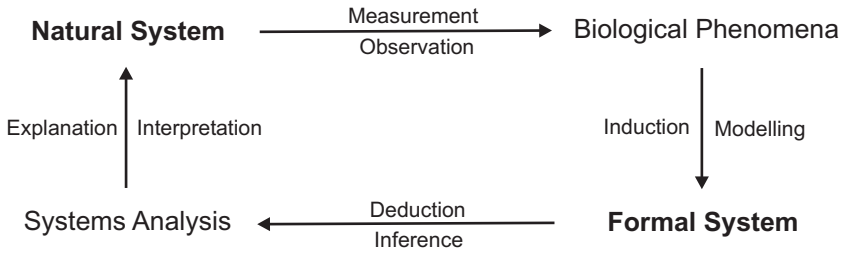


Fig. 4.2 Systems biology: systems thinking in genomics.

obtain representations of anticipatory systems. Rosen was looking for ways to characterize molecular and genetic systems in a general way and quite independently of their physical or chemical constitution. His (M,R)-systems, which we reviewed here, are unlikely to become a methodology that is *useful* to biologists. However, they serve as an example of a mathematical study of basic biological principles. His conceptual framework arose from a criticism of the transfer of principles of Newtonian physics to biology. It is in this context that his work deserves renewed interest in the post-genome era of biology and bioinformatics.

One of the challenges for the emerging field of systems biology is then to link abstract mathematical models, like for example (M,R)-systems, to specific current problems of genomics. An important difference to the 1960s is the availability of three types of gene expression data at different levels: genome level (sequences), transcriptome level (microarrays) and proteome level (mass spectroscopy, gel techniques). In particular with microarrays we can now conduct time course experiments, generating data suitable for time-series analysis. With the shift of focus from molecular characterization to an understanding of functional activity in genomics, systems biology can provide us with methodologies to study the organization and dynamics of complex multivariable genetic pathways. What are then the conditions for systems biology to succeed?

Mihajlo Mesarović wrote in 1968 that "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." His advice for the biologists was that progress could be made by more direct and stronger interactions 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 bio-

physical 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." [41].

5

Bioinformatics

New technology means that we now can observe biological systems at the molecular level. As a result, molecular biology currently witnesses a shift of focus from molecular characterisation to the understanding of functional activity. The two central questions are "What are the genes' biological function?" and "How do genes and/or proteins interact?". In the past single genes were studied but now genomics researchers measure the activity levels of thousands of genes at the same time. With microarray technology it is possible to identify interrelationships between groups of genes ("gene function") and gene interactions, for both comparative studies and over time. 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. Formal mathematical modelling of these interactions will therefore play an increasingly important role and with the emphasis on linkages, and relationships between proteins and genes, many problems become conceptual rather than just empirical.

For those interested in systems and control theory, the processes considered in the modern life-sciences provide numerous challenges and opportunities. In response to the challenge, systems and control theory are returning to their roots in cybernetics, concerned with the mathematical modelling of natural and physical systems in general. In bioinformatics, we require strategies to train scientists and engineers for a good understanding of ways and means

of how to encode the natural world into 'good' formal structures. The main technical challenges are

Dimensionality: Very large number of variables (genes).

Uncertainty: Measurement noise, imprecision, missing data and outliers. Often only a very small number of samples.

Complexity: Processes highly interconnected, hierarchical, non-linear, time-variant, adaptive.

Observability: Arrays provide only limited and indirect view of gene expression. Information from the transcriptome and proteome level need to be fused in an integrative approach. Studying the dynamic response of a cell, the system is observed in "closed loop".

5.1 BIOINFORMATICS AND SYSTEMS BIOLOGY

The area of bioinformatics has provided an important service to biologists; helping them to visualise molecular structures, analyse sequences, store and manipulate data and information. These activities will continue to be an essential part of bioinformatics, developing working methodologies and tools for biologists. However, in order to directly contribute towards a deeper understanding of the biology, bioinformatics has to establish a conceptual framework for the formal representation of interrelationships and interactions between genes or proteins. At this point the application of systems theory to genomics emerges as *systems biology*. Systems biology aims at a system-level understanding of genetic pathways by investigating interrelationships (i.e. organisation or structure) and interactions (i.e. dynamics or behaviour) of genes and proteins.

In my view, the biggest challenge of bioinformatics is not the volume of data, as commonly stated, but the formal representation of knowledge. Knowledge is the result of an exploratory, recursive process – the comprehension of information which in turn is extracted from experimental data and observations. A look at biology textbooks shows that at present, biological knowledge is primarily encoded in a textual form and using diagrams. For knowledge to be useful, i.e., accessible and verifiable; it has to be formalised. The amount of knowledge stored in texts and the preference for natural language among biologists, has lead to a number of bioinformatics research projects text mining: knowledge discovery in scientific texts.

From the current research literature, we find that the areas of proteomics and genomics currently undergo changes which have a significant impact upon the area of bioinformatics: While in the past single genes were studied, with

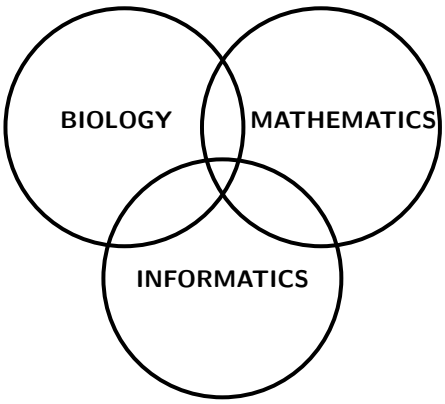


Fig. 5.1 The area of bioinformatics is an interdisciplinary effort between biology, mathematics and informatics.

microarray technology we can now measure the activity levels of thousands of genes at the same time. Secondly, proteomics research has shown that most proteins interact with several other proteins and while in the past the function of a protein was for example described by its role as a catalyst in a reaction it is increasingly appreciated that the function of a protein is appropriately described in the context of its *interactions* with other proteins. The two fundamental questions of genomics are:

"What are the genes' biological functions?"

and

"How do genes/proteins interact?"

The first question refers to *interrelationships* in sets of genes, and is effectively a problem of describing *organisation*, using *pattern recognition* techniques. The second question regarding the *behaviour* of molecular or genetic systems is appropriately dealt with in *system theory*. We refer to the combination of pattern recognition techniques and system theory as *data engineering*. In this chapter we will provide an example for such a combination.

From the developments in genomics and proteomics it becomes obvious that the challenge for bioinformatics is *not* the overused statement that the volume of data in molecular biology is increasing, but the representation of the diverse interrelationships and interactions appearing in genetic or molecular systems. The focus is shifting from molecular characterisation to an understanding of

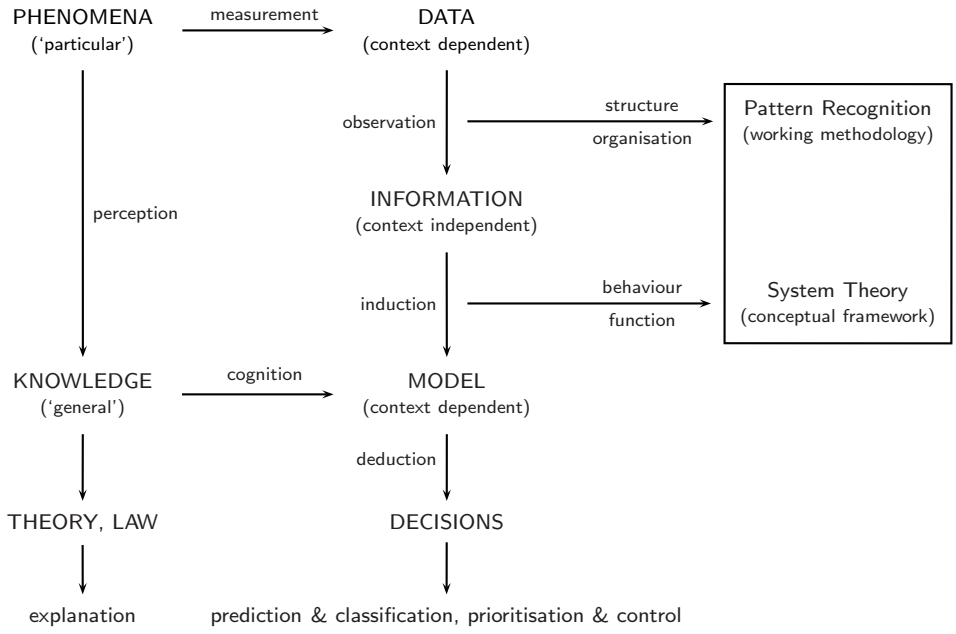


Fig. 5.2 Data engineering is the combination of pattern recognition techniques with system theory. The diagram outlines the objectives and process of data engineering applied to the life-sciences.

functional activity. As a consequence, problems in the life science will become also conceptual as well as empirical. The two key questions are:

"How do we manage the complexity of these systems?"

and subsequently

"How can we be precise about uncertainty?"

... to allow reasoning in the presence of uncertainty. Note that we suggest that complexity is to be considered a cause of uncertainty. The answer to these questions lies in the way we encode relationships, specifically *interrelationships* and *interactions*. The approach taken in this text, is summarised in Figure 5.3.

The purpose of mathematical modelling in the life sciences should not be to replicate the biologists work on paper or in a computer; mathematical models can only complement the biologists work and the role of systems theory is to provide a way of thinking, to help design experiments, to decide which

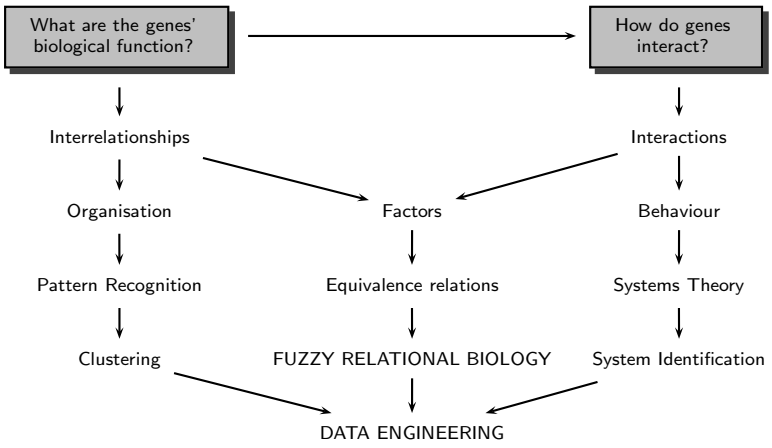


Fig. 5.3 Systems biology: Summary of the approach taken.

variables to measure and why. To achieve this perspective of mathematical modelling we can learn from the engineering sciences. Over the last century, engineers have found themselves in a situation similar what molecular biologists experience now: the systems they investigate are complex in the sense that observations cannot be simply analysed using 'common sense' or 'intuition'. In many cases, the dimensionality, non-linearity and dynamics of the system are incomprehensible to common reasoning, regardless of the expertise or experience of the investigator.

Engineers have learned to translate the given practical problem into a set of (usually random- or state-) variables and then to use a conceptual framework (probability theory, control theory) to investigate relationships between variables using a mathematical or formal system. These relationships, encoded by the mathematical model, will reflect the scientists' view of the biological problems. From systems theory we know that there is no single valid model structure and set of parameters to appropriately describe a set of data and hence the conclusion that in future we are going to see interesting debates among biologist on the interpretation of models intended to represent genetic pathways.

To illustrate the use of mathematical models, consider the experiment of rolling a dice. Asking for the likelihood or probability of any specific number to show up, we "know" from physical symmetry that each side has an equal chance to appear and hence the probability is one over six. However, this inference already assumed a model and assumptions. We assumed phys-

ical symmetry and thereby generalised the problem - ignored the homemade dice I hold in my hands and instead consider 'some' dice. This process of abstraction through generalisation *is* mathematical modelling. Mathematical concepts are abstract mental objects which, through the phenomena they are to represent, can be experienced. Whether we like it or not, we all use mathematical modelling, we all go shopping. Numbers, for example, are such mental construction and although arithmetic doesn't make the shopping easier, it helps us making the right decisions on the way to the till.

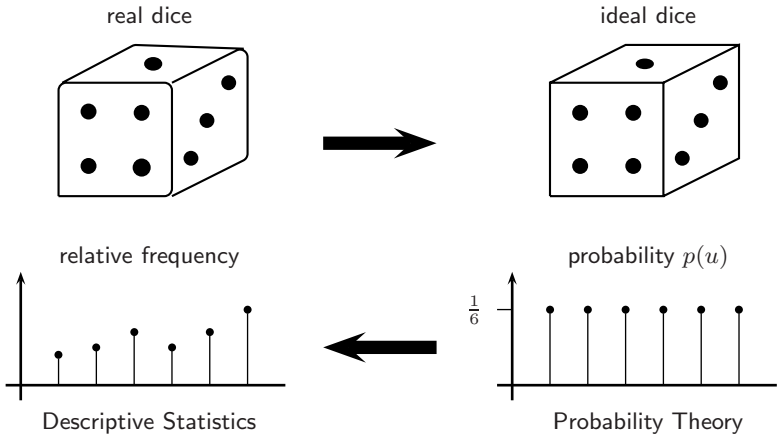


Fig. 5.4 Mathematical modelling as an abstraction. A specific real-world process is generalized by using a conceptual framework, such as probability theory, to represent it. To verify a model, a real-world interface, such as descriptive statistics, is used to validate a concept with data.

As shown in Figure 5.4, studying the dice we 'translate' the real world problem into a conceptual framework - probability theory in this case. Once a model is obtained, we validate it with experimental data. For a homemade dice we may then find that the distribution favors larger numbers. This unexpected result forces us to rethink our model and its assumptions. We may find that drilling holes into the dice, the sides with small numbers become heavier than its opposite side (both sides add up to one). We can then use basic physical principles to explain the result, dismiss the model or revise the experimental set-up.

The mathematical model requires assumption on the actual process and in the formal system. The main elements of the formal system are the *probability distribution*

$$\begin{aligned} p\colon S &\to [0,1] \\ s &\mapsto p(s) \end{aligned}$$

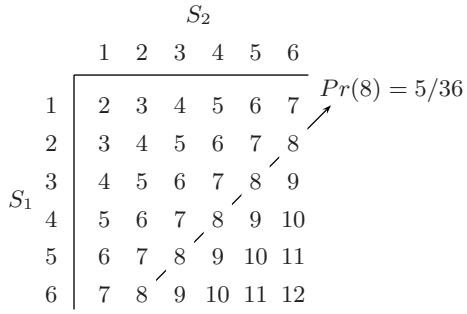


Fig. 5.5 Sample space for an experiment involving two ideal dice.

where s denotes an elementary outcome of the experiment and $p(s)$ is its likelihood. For more general *events*, represented by subsets of sample space $S = \{s\}$, the probability measure, $Pr(A)$, of the event $A \subset S$ should satisfy three further conditions:

$$\begin{aligned} Pr(S) &= 1 \\ Pr(A) &\geq 0 \\ Pr(A \cup B) &= Pr(A) + Pr(B) \end{aligned}$$

where $Pr(\{s\}) \doteq p(s)$. To investigate the result of rolling two dice, let use introduce a *random variable* f which is to describe the sum of the two numbers showing up:

$$\begin{aligned} f\colon \Omega \to R_f &= \{2,3,4,5,6,7,8,9,10,11,12\} \\ \omega \mapsto f(\omega) &= s_1 + s_2 \ . \end{aligned}$$

We notice that a random variable is neither random nor variable - it is a (real-valued) mapping with domain Ω and range R_f. The range R_f of the mapping f are all possible sums of any two numbers showing up. In general, a random variable is a mapping from the sample space of possible outcomes to a space of real numbers. The elements of the sample space can be mathematical or material objects and $f(\omega)$ describes some observable characteristic. The sample space Ω in this case is the product space $S_1 \times S_2$ of the two dice:

$$\Omega = \{(s_1, s_2)\colon s_1 = 1,2,...,6 \ s_2 = 1,2,...,6\}$$

where $\omega \doteq (s_1, s_2) \in \Omega$. The event of the sum of any two numbers showing up being equal to three, can be represented as the subset $\{(1,2),(2,1)\}$ of Ω or the set $\{3\}$ of R_f.

As illustrated in Figure 5.5, there are 36 possible elementary outcomes with equal probability $p(\omega) = 1/36$. The probability that the sum of the two dice is equal to eight than becomes $Pr(8) = 5/36$. To investigate the *concept* "the outcome is an even number", its extension is the singleton set $\{2, 4, 6, 8, 10, 12\}$. With the probability law $p(\omega)$ defined on Ω, the probability of an event A in R_f is determined by relating the set $A \subset R_f$ to $p(\omega)$ on Ω. This is achieved using the inverse mapping $f^{-1}\colon A \mapsto f^{-1}(A) = \{\omega\colon f(\omega) \in A\}$:

$$\begin{aligned} Pr(A) &= Pr_\Omega \circ f^{-1}(A) = Pr_\Omega\left(f^{-1}(A)\right) \\ &= Pr_\Omega\{\omega\colon f(\omega) \in A\} \\ &= \sum_{f \in f^{-1}(A)} p(\omega)\ . \end{aligned}$$

Figure 5.6 shows the probability distribution over R_f and a possible scenario using 'homecrafted' dice.

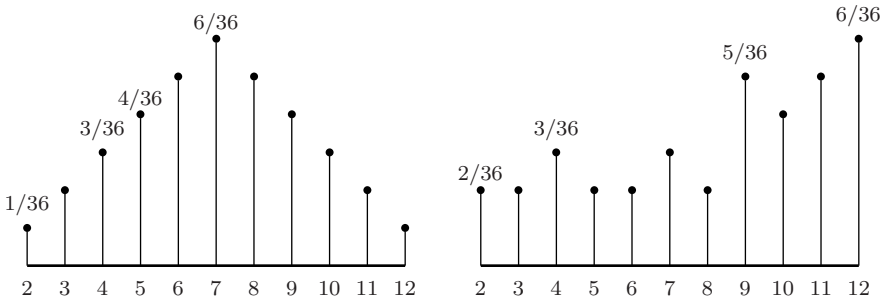


Fig. 5.6 Probability distribution for the sum of numbers showing when rolling a pair of ideal dice (left). Sample distribution from using a homemade pair of dice (right).

I believe it was the molecular biologist Jacques Monod who said 'we might prefer to ignore philosophical questions but we cannot avoid them'. For me, the most exciting question of "how nature works" is necessarily connected to the question of "how do we know?" In the post-genome era, this epistemological problem is related to the validity of the many mathematical or computational models that have been proposed. Depending on who you ask, say computer scientists, a control engineer, a "Bayesian" etc., each will have a different approach to establishing a model and even if at a formal level the result is equivalent, the semantics usually differ.

In engineering, a mathematical model is primarily expected to provide numerically accurate predictions; the internal structure and the interpretation of the model are often only of secondary interest. In life sciences, however, the purpose of models is not just replicating the external or observable behaviour

of a molecular or genetic system but also helping the biologist explain the underlying mechanisms. The semantics of the (mathematical) model should therefore have biological significance.

I personally have my doubts about what we can achieve with formal systems. Artificial Neural Networks and Genetic Algorithms are only two examples of computational techniques for which nature provided the inspiration. Their successful application to modelling and optimisation in many non-biological fields has established them (but without any relevant reference to the biological ancestry). No doubt the knowledge-transfer from nature to, for example, the engineering sciences, has been successful but we should be careful and avoid the reverse approach. I am very suspicious of suggestions that nature performs "computations" or optimises a "cost function". We should never forget that although we may find reasonable representations of natural systems, philosophers like Arthur Schopenhauer and scientists like Robert Rosen demonstrated long ago that this does not imply knowledge of "what the things are in themselves".

In the present book, we are interested in the development of a mathematical framework for modelling genome expression and regulation. Based on Schopenhauer's unsurpassed discussion of causation, causal entailment in natural systems is identified as the principle of explanation of change in the realm of matter. Causation is therefore a relationship, not between components, but between changes of states of system. I subsequently view genome expression (formerly known as 'gene expression') as a dynamic process and use methodologies developed within the areas of systems and control theory. Norbert Wiener pointed out the importance of what we now call Systems Biology when he introduced the area of Cybernetics in 1948. Wiener's work was followed by intensive research into the mathematical foundations of systems and control theory. These developments have been accompanied by numerous applications of the theory in engineering but generally failed to impress molecular biologists.

In the 1970's, Francois Jacob and Jacques Monod investigated regulatory proteins and the interactions of allosteric enzymes. They introduced a distinction between 'structural genes' (coding for proteins) and 'regulatory genes', which regulate the rate at which structural genes are transcribed. This control of the rate of synthesis of proteins gave the first indication of such processes being most appropriately viewed as dynamic systems driven by a multitude of factors and feedback regulated. Indeed, negative feedback is used in all cells and in metabolic pathways in particular. Control of such processes is achieved through regulatory enzymes that respond to effector concentrations by increase or decrease in their reaction rates. Despite their insights, biologists prefer to describe cellular processes in terms of material objects and their spatial relationships. To represent such knowledge diagrams, pictures,

and natural language are used in analysis and inference. However as Jacob and Monod have already suggested, such "Lego-style" modelling may not be the optimal approach to explain gene expression and regulation. Instead of trying to identify 'genes' as causal agents for some function, role, change in phenotype or the cellular response of proteins, we should identify these observations with sequences of events. In other words, instead of looking for a 'gene' (whatever that may mean) that is the reason, explanation or cause of some phenomenon we should seek an explanation in the dynamics (sequences of events ordered by time) that led to it. Molecular biology has focussed on the physico-chemical characterization of "parts and components" but with the emergence of new "post-genome technologies", a shift of focus to an understanding of functional activity is taking place.

While systems and control theory provides us with precise inference and quantitative predictions, it "works" only for systems that, compared to cellular processes, are simple. Biologists on the other hand have been successful in describing complex systems using empirical means and qualitative reasoning. Our mathematical models of dynamic systems are not the long awaited solution to questions in genomics but a combined effort in the modelling process; "the way of thinking" about genome expression and regulation seems a good idea. The *phenomenological model* developed in previous chapters does not describe what 'things are in themselve' but of what we perceive and conceive about the natural system under consideration. As a mathematical model it integrates observation and measurement; experimental data and relationships.

For all the reasons mentioned above, a discussion of different mathematical or computational models of biological systems is very valuable. It is a mistake to argue or to look for a single best, true or correct model but as I tried to point out, the thinking about how we model natural systems will actually help us in understanding how nature works.

6

Symbols and Notation

$\mathbb{R}$	set of real numbers, real line.
$\to$	mapping.
$\leq$	less or equal.
$\subseteq$	subsethood.
$\in$	elementhood.
$\times, \prod$	Cartesian or direct product of sets (spaces).
$\circ$	composition.
$\vee$	disjunction.
$\wedge$	conjunction.
$\cup$	union.
$\cap$	intersection.
$\neg$	negation.
^{c}	complement.
:	"for which", "given".
$\exists$	"there exists".
$\forall$	"for all".
$\doteq$	"defined".
$\approx$	"approximately".
$\equiv$	"identical".
$\Rightarrow$	"implies", material implication.
$\therefore$	"therefore".
$\mapsto$	"maps to".
iff	"if and only if".

U	objects $u \in U$, description frame, universe of discourse.
$\mathcal{C}$	set of concepts $C \in \mathcal{C}$.
F	set of factors.
f, g	factors, $f, g \in F$.
$\mathbf{0}$	zero factor.
$(U, \mathcal{C}, F)$	description frame.
$X(f)$	state-space of f, repres. universe, denoted X for short.
$f(u)$	state, value of f at u, $f(u) \in X(f)$.
$D(f)$	set of objects $u \in U$ for which f is relevant.
$V(u)$	set of factors $f \in V$ relevant to u.
R	relation.
$\widetilde{R}$	fuzzy relation.
$\widetilde{A}$	fuzzy set.
$\widetilde{A}(u), \mu_{\widetilde{A}}(u)$	membership of u in $\widetilde{A}$.
E	equivalence relation.
$\widetilde{E}$	fuzzy equivalence relation, similarity relation.
$\mathcal{F}(U)$	the set of all fuzzy sets in U.
$\mathcal{P}(U)$	the set of all crisp sets in U (power set).
$\mu_{\widetilde{A}}(\cdot)$	membership function of fuzzy set $\widetilde{A}$.
$\tilde{f}$	fuzzy mapping.
$\mathcal{G}$	fuzzy graph.
$T(\cdot)$	triangular norm, T-norm.
$S(\cdot)$	triangular co-norm, S-norm.
$\uparrow_g^f$	cylindrical extension.
$\mathbf{x}$	fuzzy or random variable.
$d(\cdot)$	distance, metric.
$[u]$	equivalence class.
U/E	quotient set.
$\emptyset$	empty set.
$\prec$	partial order.
$\lvert \cdot \rvert$	absolute value.
ε	error bound, tolerance.
$Pr(\cdot)$	probability measure.
$F(\cdot)$	cumulative probability.
$\widetilde{A}$	extension of C in U.
$f(\widetilde{A})$	(repres.) ext. of C in $X(f)$, $f(\widetilde{A}) \in \mathcal{F}(X(f))$.
$f^{-1}(\widetilde{B}(f))$	feedback extension of C w.r.t f.
Λ	measure of coincidence.
$\widetilde{A}[G]$	G-envelope, G-feedback extension of $\widetilde{A}$.
ρ	natural mapping.

ϕ	embedding.
$\mathcal{L}(f, g)$	measure of linkage between factors f and g.
η	general measure, count.
(U, E_f)	approximation space.
$A - B = A \cap B^c$	set difference.
$\mu_A^r(u)$	rough set membership function of A.
$\mu_{E^*(\tilde{A})}([u]_f)$	membership function of a rough fuzzy set.
$E_*(A), E^*(A)$	lower (upper) approximation of A.
$\mathcal{A}(A)$	accuracy of approximation of A in U by E_* and E^*.
$\#(B)$	cardinality of (finite) set B.
Bel	belief function.
Pl	plausibility function.
$m\colon \mathcal{P}(U) \to [0, 1]$	basic probability assignment (mass distribution).

References

1. Ashby, W.R. : *An Introduction to Cybernetics.* Chapman and Hall, London 1956. Internet (1999): `http://pcp.vub.ac.be/books/IntroCyb.pdf`

2. Bailey, J.E. : *Mathematical Modelling and Analysis in Biochemical Engineering: Past Accomplishments and Future Opportunities.* Biotechnol. Prog., Vol. 14, pp. 8–20, 1998.

3. Bailey, J.E. : *Lessons from metabolic engineering for functional genomics and drug discovery.* Nature Biotechnology, Vol. 17, pp. 616–618, July 1999.

4. Bertalanffy, L. : *General Systems Theory: Foundations, development, applications.* Harmondsworth 1968, Penguin, 1973.

5. Birkhoff, G. and Von Neumann, J. : *The Logic of Quantum Mechanics.* Annals of Mathematics, Vol. 37, No. 4, pp. 823–843, October 1936.

6. Bohm, D. : *Wholeness and the implicit order.* Routledge, 1980.

7. Boltzmann, L. : *On a thesis of Schopenhauer's.* Populäre Schriften, Essay 22. Delivered to the Vienna Philosophical Society, 21 January 1905. In 'Theoretical Physics and Philosophical Problems', edited by B. McGuiness, D.Reidel Publishing Company 1974.

8. Brown, T.A. : *Genomes.* BIOS Scientific Publishers, 1999.

9. Brown, M.P.S. et al. : *Knowledge-based analysis of microarray gene expression data by using support vector machines.* Proc. Natl. Acad. Sci. USA (PNAS), Vol. 97, pp. 262–267, January 2000.

10. Casti, J.L. : *Linear Metabolism-Repair Systems.* Int. J. General Systems, Vol. 14, pp. 143–167, 1988.

11. Casti, J.L. : *The Theory of Metabolism-Repair Systems.* Applied Mathematics and Computation, Vol. 28, pp. 113–154, 1988.

12. Casti, J.L. : *Reality Rules: Vol I + II.* John Wiley & Sons, Inc., New York, 1992.

13. Corning, P.A. and Kline, S.J. : *Thermodynamics, Information and Life revisited.* Syst. Res., Vol. 15, pp. 453–482, 1998.

14. Edward, J.S. and Palsson, B.O. : *How will bioinformatics influence metabolic engineering.* Biotechnology and Bioengineering, Vol. 58, Nos. 2 & 33, April 20/May 5, pp. 162–169, 1998.

15. Einstein, A. : *Relativity: The Special and the General Theory*, Three Rivers Press, New York, 1961.

16. Eisen, M.B. et al. : *Cluster analysis and display of genome-wide expression patterns.* Proc. Natl. Acad. Sci. USA (PNAS), Vol. 95, pp. 14863–14868, December 1998.

17. D'Haseleer, P. et al. : *Linear Modelling of mRNA Expression Levels During CNS Development and Injury.* Pacific Symposium on Biocomputing, 4:41–52 (1999).

18. Dubois, D. and Prade, H. : *Rough Fuzzy Sets and Fuzzy Rough Sets.* Int. J. General Systems, Vol. 17, pp. 191–209.

19. Farnum, N.R. and Stanton, L.W." : *Quantitative Forecasting Methods.* PWS-Kent Publishing Company, 1989.

20. Glaserfeld, von E. : *Radical Constructivism.* The Falmer Press, 1995.

21. Heinrich, R. and Schuster, S. : *The Regulation of Cellular Systems.* Chapman & Hall, 1998.

22. Höhle, U. : *Quotients with respect to similarity relations.* Fuzzy Sets and Systems, Vol. 27, pp. 31–44, 1988.

23. Höhle, U. and Stout, L.N. : *Foundations of Fuzzy Sets.* Fuzzy Sets and Systems, Vol. 40, pp. 257–296, 1991.

24. Höhle, U. : *On the Fundamentals of Fuzzy Set Theory*. Journal of Mathematical Analysis and Applications, Vol. 201, pp. 786–826, 1996.

25. Hood, L. (2000): *What is Systems Biology?* Institute for Systems Biology, Seattle, Washington, USA, See `http://www.systemsbiology.org`

26. Huyen, M.A. and Bork, P. : *Measuring Genome Evolution.* Proc. Natl. Acad. Sci. USA, Vol. 95, pp. 58–5856, May 1998.

27. IMDS project, TU Delft : *Intelligent Molecular Diagnostic Systems.* `http://www.ph.tn.tudelft.nl/~young/DIOC/IMDS.html`.

28. Jacob, F. and Monod, J. : *Genetic regulatory mechanisms in the synthesis of proteins.* J. Mol. Biol., Vol. 3, pp. 318–356, 1960.

29. Kauffman, S.A. : *The Origins of Order.* Oxford University Press, 1993.

30. Kitano, H. (2000): *Perspectives on Systems Biology.* New Generation Computing, Vol. 18, pp. 199–216. See also `http://www.systems-biology.org/`

31. Kruse, R. et al. : *Uncertainty and Vagueness in Knowledge-Based Systems.* Springer Verlag, 1991.

32. Kruse, R. and Gebhardt, J. and Klawonn, F. : *Foundations of Fuzzy Systems.* John Wiley, 1994.

33. Klir, G.J. : *Facets of Systems Science.* Plenum Press, 1991.

34. Klir, G.J. and Yuan, B. eds. : *Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems: Selected Papers by Lotfi Zadeh.* World Scientific, 1996, River Edge, New Jersey.

35. Kyrpides, N.C. : *Genomes Online Database (GOLD 1.0).* Bioinformatics, Vol. 15, No. 9, 1999, pp. 773–774.

36. Lawvere, S. and Schanuel, S. : *Conceptual Mathematics.* Cambridge University Press, 1997.

37. Li, H.X and Yen, V.C. : *Fuzzy Sets and Fuzzy Decision Making.* CRC Press, 1995.

38. Magee, B. : *The Philosophy of Schopenhauer*. Clarendon Press, 1997.

39. Marcotte, E.M. et al. : *A combined algorithm for genome-wide prediction of protein function.* Nature, Vol. 402, No. 4, pp. 83–86, 1999.

40. Mesarovic, M.D. et al. : *System Theory and Biology*. Springer, 1968.

41. Mesarović, M.D. (1968): *Systems Theory and Biology - View of a Theoretician.* pp. 59–87, in [40].

42. Mesarović, M.D. and Takahara, Y. : *General Systems Theory: Mathematical Foundations.* Academic Press, New York, 1975.

43. Mesarović, M.D. and Takahara, Y. : *Abstract Systems Theory*. Springer-Verlag, New York, 1988.

44. Miller, J.G. : *Living Systems*. University of Colorado Press, 1995.

45. Monod, J. : *Le hasard et la nécessité*.('Chance and Necessity'), Seuil, Paris, 1970.

46. Moore, W. : *Schrödinger: Life and Thought*. Cambridge University Press, 1989.

47. Muir, A. : *Holism and Reductionism are Compatible*. In *Against Biological Determinism*, Rose, S. (ed.), Allison & Busby, 1982.

48. Oliver, S.G. et al. : *From DNA sequence to biological function*. Nature, 379, pp. 597–600, 1996.

49. Oliver, S.G. : *Guilt-by-association goes global*. Nature, Vol. 403, pp. 601–603, 10th February 2000.

50. Palsson, B.O. : *What lies beyond bioinformatics*? Nature Biotechnology, Vol. 15, pp. 3–4, 1997.

51. Poincaré, H. : *Science and Hypothesis*. Dover Publications Inc., New York, 1952.

52. Principia Cybernetica Web (2001), `http://pcp.vub.ac.be/DEFAULT.html`

53. Paton, N. et al. : *Conceptual Modelling of Genomic Information*. To appear in Bioinformatics, 2000. http://img.cs.man.ac.uk/gims.

54. Pawlak, Z. : *Rough Sets*. Int. J. of Computing and Information Sciences, Vol. 11, No. 5, 1982, pp. 341–356.

55. Pawlak, Z. : *Rough Classification*. Int. J. Man-Machine Studies (1984), **20**, pp. 469–483.

56. Pawlak, Z. et al. : *Rough sets: probabilistic versus deterministic approach*. Int. J. Man-Machine Studies (1988) **29**, pp. 81–95.

57. Pedrycz, W. : *Fuzzy Control and Fuzzy Systems*. Research Studies Press, 1992.

58. Rosen, R. : *The representation of biological systems from the standpoint of the theory of categories*. Bulletin of Mathematical Biophysics, Vol. 20, 317–341, 1958.

59. Rosen, R. : *Fundamentals of Measurement and Representation of Natural Systems*. North-Holland, 1978.

60. Rosen, R. : *Anticipatory Systems*. Pergamon Press, 1985.

61. Rosen, R. : *Life Itself*. Columbia University Press, 1991.

62. Rosen, R. : *Essays on Life Itself*. Columbia University Press, 2000.

63. Russell, B. (1948): *Human Knowledge: Its scope and limits*. Routledge, London, 1992.

64. Schopenhauer, A. : *The World as Will and Representation*. Vol. 1, Dover Publications, 1967.

65. Schopenhauer, A. : *On the Fourfold Root of the Principle of Sufficient Reason*. Open Court Publishing Company, 1974.

66. Schrödinger, E. : *What is Life?*. 1944, Cambridge University Press 1992 with *Mind and Matter* and *Autobiographical Sketches*.

67. Schrödinger, E. : *My view of the world*. Cambridge University Press, 1964.

68. Schweizer, B. and Sklar, A. : *Probabilistic Metric Spaces*. North-Holland, 1983.

69. Shafer, G. : *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

70. Somogyi, R. and Sniegoski, C.A. : *Modelling the complexity of genetic networks*. Complexity 1 (6): 45–63, 1996.

71. Stephanopoulos, G.N. et al. : *Metabolic Engineering*. Academic Press, 1999.

72. Varma, A. and Palsson, B.O. : *Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use*. Bio/Technology, Vol. 12, October 1999, pp. 994–998.

73. Wang, P.-Z. : *A Factor Spaces Approach to Knowledge Representation*. Fuzzy Sets and Systems, Vol. 36 (1990), pp. 113–124.

74. Wang, P.-Z. : *Knowledge Acquisition by Random Sets*. International Journal of Intelligent Systems, Vol. 11 (1996), pp. 113–147.

75. Waterman, T.H. (1968): *Systems Theory and Biology - View of a Biologist*. Page 1–37, in [40].

76. Wiener, N. : *Cybernetics: Control and Communication in the Animal and the Machines*. MIT Press, Cambridge Massachusetts, 1948.

77. Wolkenhauer, O. : *Possibility Theory with Applications to Data Analysis*. Research Studies Press, 1998.

78. Wolkenhauer, O. : *Data Engineering: Fuzzy Mathematics in Systems Theory and Data Analysis*. John Wiley & Sons, New York 2001.

79. Wolkenhauer, O. : *Fuzzy Relational Biology: A Factor-Space Approach to Genome Analysis*. Computation in Cells: Proceedings of an EPSRC Emerging Computing Paradigms Workshop. H.Bolouri, R.C.Paton (eds.), pp. 53–58.

80. Wolkenhauer, O. : *Systems Biology: The reincarnation of systems theory applied in biology?* Briefings in Bioinformatics, Vol. 2, No. 3, 2001.

81. Yen, J. : *Generalising Dempster-Shafer theory to fuzzy sets.* IEEE Transactions on Systems, Man, and Cybernetics 20 (3), 1990, pp. 559–570.

82. Zadeh, L. : *Similarity Relations and Fuzzy Orderings*. Information Sciences, Vol. 3, pp. 177–200, 1971.

83. Ziarko, W.P. : *The Discovery, Analysis, and Representation of Data Dependencies in Databases*. In Piatesky-Shapiro, G. and Frawley, W.J. (eds.): *Knowledge Discovery in Databases*, AAAI Press/MIT Press, 1991, pp. 177–195.

84. Ziarko, W.P. (ed.) : *Rough Sets, Fuzzy Sets and Knowledge Discovery*. Springer Verlag, 1994.

85. Zweiger, G. : *Knowledge discovery in gene-expression microarray data: mining the information output of the genome*. Trends in Biotechnology, Vol. 17, pp. 429–436, November 1999.