Level 1

Genomics, Metabolomics, and *Distributed Networks*

There are very different kinds of subsystems in a living organism. To reach the level of *systemic biology,* one needs to be able to combine different model structures. The great promises have encouraged the researchers already for a long time to try and construct combined models (for example, see [85]).

When constructing models, it is first necessary to capture the essence of a domain field in mathematical structures. This is a delicate challenge, and domain-area expertise is needed. However, it seems that when one follows the neocybernetic guidelines, homogeneous representations can be defined for different kinds of systems. In this chapter, appropriate representations for information are first defined, starting from concrete low-level models, and abstracting them towards general model structures. Further analyses in subsequent chapters are based exclusively on these data representations.

1.1 Experiences with "artificial cells"

Information of the biological processes has increased immensely: Capability of reading the genetic code, and new ways of gaining information of the genetic activities (like *Chromatin Immunoprecipitation* or ChIP technique, see [74]) has delivered us large amounts of data. It has been assumed that being capable of deciphering the genetic code is enough to reach the higher level of understanding of what life is. Perhaps some day all dependencies between phenomena and control structures within a cell have been found. Is this not the ultimate goal?

Unfortunately, this is just one step towards capturing the essence of life processes, and the goal will *never be reached this way.*

What is the contribution an engineer having no background in biology can have, giving advice to domain area experts? — Counterintuitively, the engineer's contribution can offer wider views here. Systems thinking is universal, and, experiences from other fields can be exploited. Perhaps the same deadlocks can be avoided?

Understanding of complex systems is the challenge also in industrial automation systems. Despite the detailed system models, computers and simulators, the behaviors and qualitative properties of the overall system are becoming more and more difficult to understand. The systems are becoming like *artificial cells* themselves:

> Industrial plants also have *metabolism,* raw materials being exhausted and others being produced. Originally, the production can be far from optimum, but as soon as dependencies among variables are recognized, they can be used for constructing new feedback structures to implement more efficient and robust production. However, as the complexity of control structures cumulates, the system-level properties cannot any more be easily seen — even though all individual control structures are explicitly known, indeed, even though they have been explicitly designed and optimized.

In both cases, natural and man-made cells alike, it turns out that the goal of "evolution" is overall efficiency of production, no matter whether it is humans that are acting as agents for development or not. This can be reached by implementing mechanisms for reaching best possible production conditions; and this system integrity needs to be maintained without collapses. To maintain such balance, the system has to respond appropriately to the spectrum of disturbances coming from the environment.

Mastering huge amounts of data and finding "holistic understanding" out from it — this is the common goal in both cases, in artificial and natural cells alike. And, indeed, here it is the engineering tools and intuitions that can be exploited to find new kinds of approaches for attacking the complexity.

So, how to see the forest for the trees — how to see the cell metabolism for the chemical reactions, or, further, how to see the organ functions for the cellular phenomena? First, the details of the systems need to be understood and captured in data.

1.2 Modeling cellular processes

When aiming towards truly adapting systems, the model structures should not be fixed beforehand. It has to be assumed that there is minimum number of preprogramming, and the final structures have to be extracted directly from the observation data. One needs strong consistent frameworks where the observations can be interpreted. The hints of structure have to be coded in the data, and the mathematical machinery has to be capable of exploiting these hints. Indeed, this is a very ambitious goal.

1.2.1 From formulas to behaviors

Traditionally, modeling of complex systems is like search for the philosopher's stone: One tries to find the magical formula that explains all behaviors. Indeed — all systems, large and small, are assumed to be governed by underlying

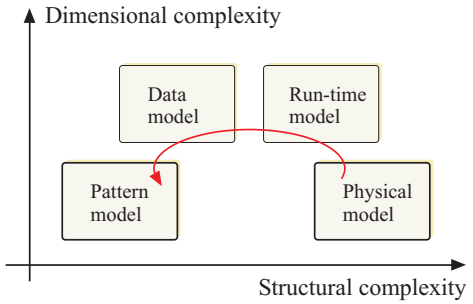

Figure 1.1: Reaching structural and dimensional simplicity at the same time

formulas, the complexity typically being manifested in nonlinearity; this kind of thinking is reflected also in today's approaches in complexity theory. However, in the case of truly large systems this objective becomes obsolete and absurd: As the systems are distributed, their behaviors cannot be compressed into some compact kernel.

Traditionally one tries to find the simplest possible formula that describes the system as isolated from its environment and other systems. The system alone is thought to represent itself in the most accurate way. But natural systems are never alone. In line with the neocybernetic assumption of environment-orientedness, it is assumed that it is the environment of the system that determines how the system is run and which of the potential behaviors become actual, selecting the subset of possible behaviors that is to be excited. In a sense, the environment carries out experimenting with the system, changing the conditions, and the system responds, finding a new balance reflecting the properties of the coupling between the system and its environment.

To unambiguously characterize the system behaviors, also the properties of its environment need to be quantified. In a complex environment, the simplicity goal cannot any more be reached as there exist a multitude of variables determining the environment. One is facing a problem of complexity exploding in two ends: Instead of simplicity, there is structural complexity in the form of the system model, and there is the dimensional complexity in the form of environmental variables.

However, as shown in Fig. 1.1, the structural complexity of the model can be ripped off by letting the complex behaviors be "interpreted" by the environment. When the environment is seen as a "simulator", and when one records the resulting behaviors, one has only homogeneous, bare numbers left. There is a high number of this kind of structureless data; assuming that the data is collected in an appropriate way, the relevant behaviors are present in that data, yet in a highly redundant form. The remaining task is that of detecting the patterns, finding the compressed representation of the information buried in the data — if this can be accomplished, one has a representation of the system that is simple in terms of structure and dimension at the same time.

How is such simplification possible, and how could it ever be done without human control and specific domain-area knowledge? First, the structural complexity in the form of nonlinearities can be changed into dimensional complexity when different kinds of nonlinearity prototypes are included in the data. Among the multitude of simple nonlinear features, the original nonlinearity can be ap-

proximated as a weighted combination of simpler ones. The question what these domain-specific nonlinearity prototypes are and how they can be isolated by applying appropriate data preprocessing so that model structure itself can be kept simple and general is discussed in this chapter, concentrating on the domains that are characteristic to a biological cell. After this, when the "behavioral essence" is present in the data, the second phase becomes possible: That of compressing the internal structure in the data and crystallizing the behavioral features. Only by making some assumptions about the nature of information (see chapter 2), the compression of the high number of environmental and system-specific data is possible.

The complexity becomes parameterized; the more there are features, the more accurately the behaviors can be captured. One can question whether the original function form can be represented by a set of other functions — but from the modeling point of view, if there is no difference in behaviors, the implementation has no relevance. According to the *identity of indiscernibles* originally due to Leibniz one can even claim that, "if all attributes of items A and B are identical, A and B are the same". Here data represents structure, the observations being assumed to capture the identity of the object, relevance of phenomena is defined in terms of visibility in data: As interactions make a difference, everything that is of relevance must be observable. This has only pragmatic motivation — but, later, closer look in these issues is taken: After all, the natural systems also only see the data available in their environments, and they try to tackle with it.

High dimensionality is easier to tackle with than different kinds of complicated structures — it turns out that the same tools are applicable to all systems. When the complexity has been transformed into high dimensionality, one can characterize the modeling task as being a *search problem*: When determining the model, or the patterns characterizing the data, one is facing a technical problem of finding the location in the parameter/variable space that corresponds to the observed behaviors. As the dimension grows, the search space grows exponentially. However, in the appropriate mathematical framework this search can be implemented efficiently in a *parallel* form. In (linear) vector spaces, it is multivariate statistical methods based on linear algebra that turn out to be efficient tools (see chapter 2).

The delicate relationship between the system and its environment is studied later closer; here it is enough to observe that all relevant nuances are represented in the data. Data selection essentially affects the modeling results, and selecting reasonable data and preprocessing it appropriately is a key question. One needs to find a coding where the system state can be captured in data; optimality of the representation needs not yet be worried about. To reach this, something needs to be known about the system structure — here it is assumed that what one is looking for is representations of *networks.*

1.2.2 Approaches to networks

It seems that an appropriate framework for studying the spectrum of the distributed cellular processes is that of networks, consisting of more or less independent interacting actors. The genetic system constitutes a network, where genes regulate each other, and also the metabolic system can be seen as a net-

work among chemicals. And other complex systems, too — like ecosystems — are networks between individuals. How to model such networks? There exist dozens of alternative approaches — let us study some examples from opposite ends of the continuum.

Graph theory

The traditional approach to modeling networks, dating back to Leonhard Euler, was graph theoretic: The connections between nodes are assumed to be "crisp" — either there is an arc or there is not. Causal structures can be modeled using directed graphs with unidirectional arcs. This is also the traditional approach, for example, when representing control relations among genes, or when representing metabolic cycles.

However, such graphs are *descriptive* as models — easy to grasp but difficult to apply. There is the same problem as there is with all qualitative models: Just knowing that there is some connection is not enough. Altering the threshold level, the network easily changes from sparsely connected to more or less fully connected. And as is known in control engineering, the numeric values in the feedback structures essentially determine the properties of the whole system.

Indeed, in real networks, there is a continuum of interaction effects: The connections are not of "all-or-nothing" type. The graph models can be extended by adding weights to the arcs, etc., but at some point it is better to rethink the structure all over.

Probabilistic networks

Probability theory offers consistent ways for defining weighted arcs. In Bayesian networks, the theory of conditional probabilities is exploited, and chains of "evidence" and resulting probabilities are expressed as tree structures. Whereas graphs suffer from weak mathematical theory, Bayesian networks benefit from strong underlying mathematics – assuming that the assumptions hold: The evidence have to be independent of each other, etc. However, the nodes in real life are often not independent of each other. There exist feedback loops and alternative paths in complex networks, making the conditional structure intractable, or at least very complex (see [62]).

Another popular model family for capturing dynamic phenomena is based on Markov models, where the state transitions are probabilistic. However, now the problems are caused by the dynamic structure: It is difficult to find the typically large number of free parameters in such models. In real life, only a too narrow view of the potential dynamics is seen to identify the parameters; experiments typically are not very *persistently exciting.* And, again, there is the basic problem: The causality structure should be explicitly resolved to reach useful models.

Neocybernetic approach

After all, practically every node in a complex network is connected to all others, either directly or through intermediate steps. The neocybernetic approach is explicitly opposite to the traditional interaction-at-a-time analyses: Now, *pan-causality* is taken as the starting point. It is assumed that, after all, all nodes are simultaneously causes and all are effects, with the exception of the explicit system inputs. The network becomes more or less fully connected.

In balance, after the transients have decayed after some disturbance, the causal effects find their balance of tensions, assuming — in the neocybernetic spirit — that the underlying interactions and feedbacks are capable of maintaining the balance. It does not matter what the details of this stabilization are as long as the balance always is finally found.

The neocybernetic model tries to avoid the above problems of traditional network models. It is a numeric model, consisting of non-crisp connections, but the numeric values of these connections are not determined applying some probabilistic hypotheses, but by observing the relations between the materialized node activities. There is no centralized control or explicit network structure assumed or solved, so that there is no need to determine for the individual interaction strengths. Indeed, as is well known, the substructures in a closed loop system cannot be distinguished — this truth is implicitly accepted in the neocybernetic framework.

If the actual causality structures and dynamics are ignored, one could wonder, what is there left of the system? The neocybernetic model is a static balance model, or actually a model over the spectrum of balances as the environmental inputs change. If the system happens to be linear, the system state is unique, being a linear function of the input. These issues will be studied in more detail later, and they will be exploited accordingly. Next it will shown that, truly, the dependencies among variables in biological networks can be assumed (locally) linear.

1.3 Case 1: Metabolic systems

As presented in [92], the original starting point in neocybernetic studies was analysis of *Hebbian neurons* — however, when modeling biological systems, such analyses do not deliver useful information. It is *proteins,* for example, that are the means for implementing the cellular and organ-specific behaviors: A more relevant framework in this case is that of *organic chemistry.* It is the metabolic processes that eventually determine the cell behaviors, being the visible manifestations of the cell character.

Study a hypothetical example reaction, where there are α reactants on the left hand side, being denoted as A_i, $1 \leq i \leq \alpha$, and the β products on the right hand side are B_j, $1 \leq j \leq \beta$:

$$a_1\mathrm{A}_1 + \cdots + a_\alpha\mathrm{A}_\alpha \quad \underset{k_\mathrm{A}}{\overset{k_\mathrm{B}}{\Leftrightarrow}} \quad b_1\mathrm{B}_1 + \cdots + b_\beta\mathrm{B}_\beta, \quad \Delta H. \tag{1.1}$$

The metabolic processes are typically reversible, so that the reaction can take

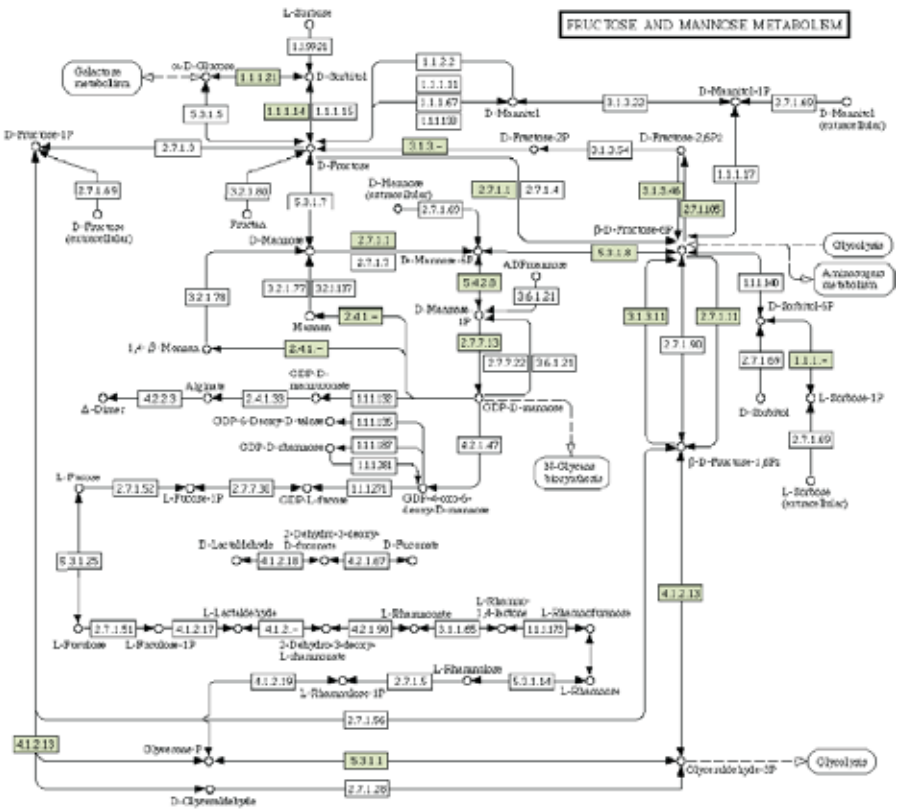

Figure 1.2: An example of metabolic pathways (courtesy of MediCel)

place in both directions (k_{B} being the reaction speed in forward and k_{A} in backward direction). Symbol ΔH denotes the change in enthalpy, or inner energy, when the reaction takes place. It needs to be recognized that it is not only chemical reactions that can be expressed using such formulas; also phase transitions, etc., can be expressed in this form.

However, chemical processes in metabolic systems can be very complex. For example, the active reaction chains in yeast when mannose production is taking place is shown in Fig. 1.2. And, what is more, such chains are just a part of the story: There exist overlapping sub-networks, and depending of the "projection", the outlook of the graph changes. If some connection is explicitly cut, for example, applying some gene kick-off technique, the results are typically not what one would expect. The cytoplasm is typically strongly buffered — there seem to exist reserve mechanisms for compensating for the disturbances.

One needs a mathematically more compact representation for chemical reactions. How to "cybernetize" chemical reaction models applying the neocybernetic principles?

1.3.1 Applying the neocybernetic guidelines

In short, the goal here is to capture the domain area semantics, or expert knowledge in distinct pieces of information, and then pack this information into a compact form that makes it possible to apply the mathematical machinery. What is this domain-area semantics to be coded, then? From the point of view, the key point is to somehow capture the balances in the system.

Information representation

The first problem is how to represent such a chemical reaction formula in a useful numeric form. It seems that a practical way to code the reactions in a mathematically applicable form is to employ vector formulation: Define a vector C containing all chemical concentrations so that all A_i and B_j are represented there among the elements. The "chemical state" can assumedly be captured in this vector.

Let us look how this vector presentation can be exploited. If the coefficients $-a_i$ and b_j from (1.1) corresponding to the chemicals are collected in the vector *Gamma*, one can express the total concentration changes in the system as

$$\Delta C = \Gamma\, x. \tag{1.2}$$

Here, x is a scalar that reveals "how much" (and in which direction) that reaction has proceeded. When there are many simultaneous reactions taking place, there are various vectors Γ_i; the weighted sum of reaction vectors Γ_i reveals the total changes in chemical contents (assuming that the vectors are compatible).

Using the above framework, metabolic systems can in principle be modeled: If one knows the rates of reactions, or the scalars x_i, the changes in the chemical contents can be determined. This idea of *invariances* within a chemical system have been widely applied for metabolic modeling; the key term here is *flux balance analysis (FBA)* (for example, see [27]). However, the rates x are not known beforehand, and, what is more, the reactions are typically not exactly known.

In many ways, the model structure (1.2) is not yet what one is looking for. The main problem there is that the flux balances only capture the *stoichiometric,* more or less *formal balance* among chemicals. It does not capture the *dynamic balance,* whether or not the reactions actually take place or not. Luckily, there exist also other ways to represent the chemical realm.

Thermodynamic balance

There is a big difference between what is *possible* and what is *probable,* that is, even though something may happen in principle, it will not actually happen. To understand the dynamic balance, the reaction mechanisms need to be studied closer.

Assume that it takes a_1 molecules of A_1, a_2 molecules of A_2, etc., according to (1.1), for one unit reaction to take place. This means that all these molecules have to be located sufficiently near to each other at some time instant for the

forward reaction to take place. The probability for one molecule to be within the required range is proportional to the number of such molecules in a volume unit; this molecular density is revealed by concentration (when the unit is mole/liter; by definition one mole always contains $6.022 \cdot 10^{23}$ particles). Assuming that the locations of the molecules are independent of each other, the probability for several of them being found within the range is proportional to the product of their concentrations. On the other hand, the reverse reaction probability is proportional to the concentrations of the right-hand-side molecules. Collected together, the rate of change for the concentration of the chemical A_1, for example, can be expressed as a difference between the backward reaction and forward reaction rates:

$$\frac{d\,C_{\mathrm{A}_1}}{d\,t} = -k_{\mathrm{B}} C_{\mathrm{A}_1}^{a_1} \cdots C_{\mathrm{A}_\alpha}^{a_\alpha} + k_{\mathrm{A}} C_{\mathrm{B}_1}^{b_1} \cdots C_{\mathrm{B}_\beta}^{b_\beta}. \tag{1.3}$$

In equilibrium state there holds $\frac{d\,C_{\mathrm{A}_1}}{d\,t} = 0$, etc., and one can define the constant characterizing the thermodynamic equilibrium:

$$K = \frac{k_{\mathrm{B}}}{k_{\mathrm{A}}} = \frac{C_{\mathrm{B}_1}^{b_1} \cdots C_{\mathrm{B}_\beta}^{b_\beta}}{C_{\mathrm{A}_1}^{a_1} \cdots C_{\mathrm{A}_\alpha}^{a_\alpha}}. \tag{1.4}$$

Linearity objective

One of the neocybernetic objectives is that of linearity. Clearly, the expression (1.4) is far from being linear — indeed, it is purely multiplicative. It turns out that applying a purely syntactic trick, linearity of the structures can be reached: Taking logarithms on both sides there holds

$$\log K' = b_1 \log C_{\mathrm{B}_1} + \cdots + b_\beta \log C_{\mathrm{B}_\beta} - a_1 \log C_{\mathrm{A}_1} + \cdots - a_\alpha \log C_{\mathrm{A}_\alpha}. \tag{1.5}$$

To get rid of constants and logarithms, it is also possible to differentiate the expression:

$$0 = b_1 \frac{\Delta C_{\mathrm{B}_1}}{\bar{C}_{\mathrm{B}_1}} + \cdots + b_\beta \frac{\Delta C_{\mathrm{B}_\beta}}{\bar{C}_{\mathrm{B}_\beta}} - a_1 \frac{\Delta C_{\mathrm{A}_1}}{\bar{C}_{\mathrm{A}_1}} + \cdots - a_\alpha \frac{\Delta C_{\mathrm{A}_\alpha}}{\bar{C}_{\mathrm{A}_\alpha}}, \tag{1.6}$$

where the variables $\delta C_i = \Delta C_i / \bar{C}_i$ are deviations from the nominal values, divided by those nominal values, meaning that it is *relative changes* that are of interest. The differentiated model is only locally applicable, valid in the vicinity of the nominal value.

Multivariate representation

A single reaction formula can also be expressed in a linear form when the variables are appropriately selected. However, to model complex systems consisting of various reactions, the data representation needs to be extended: The differing data vectors containing different sets of variables (the reactions employing different chemicals) have to be embedded in the same vector space to make them compatible.

Assume that the vector z is a vector containing all relevant variables capturing the state of the environment and the system itself, including, for example, relative changes in all chemical concentrations. This means that the vector Γ_i representing a single reaction can contain various zeros, assuming that the corresponding chemicals are not contributing in the reaction i. If the vectors Γ_i are collected as columns in the matrix Γ, one can write the individual expressions in (1.6) in the matrix form where one row is allocated to each of the reactions:

$$0 = \Gamma^T \, \delta z, \tag{1.7}$$

or, when written out,

$$\left\{ \begin{array}{rcl} 0 & = & \Gamma_{1,1}\delta z_1 + \cdots + \Gamma_{m,1}\delta z_m \\ & \vdots & \\ 0 & = & \Gamma_{1,n}\delta z_1 + \cdots + \Gamma_{m,n}\delta z_m, \end{array} \right. \tag{1.8}$$

where n is the total number of reactions, and m is the total number of chemicals. This expression needs to be compared to flux balance analysis: Now one only needs to study levels of concentrations, not changes in them. This is indeed essential in complex chemical systems, where the energy and matter flows cannot be exactly managed. The key point to observe here is that analysis of complicated reaction networks can be avoided: No matter what has caused the observed chemical levels, only the prevailing tensions in the system are of essence. The underlying assumption is that the system is robust and redundant: Individual pathways are of no special importance as there exist various alternative routes in the network.

It turns out that reactions can in principle be characterized applying linear algebra in the space of chemical concentrations. However, in practice it is not enough to only represent the concentrations if the properties of the whole system need to be captured. What else can the vector u contain?

1.3.2 Characterizing the metabolic state

The measurement vector z needs to be further studied to make it possible to capture all *internal tensions* in metabolic systems. As it turns out, the following extensions can, for example, be implemented without ruining the linear structure among the variables:

- **Temperature.** According to the Arrhenius formula, the reaction coefficients are functions of the temperature, reactions becoming faster as the temperature rises, so that $k \propto \exp(c/T)$. This means that when this is substituted in the formulas, and when logarithms and differentiations are carried out, the model remains linear if the new variable is defined as $z_T = \Delta T/\bar{T}^2$.
- **Acidity.** The pH value of a solution is defined in terms of a nonlinear formula: $\mathrm{pH} = -\lg C_{\mathrm{H}^+}$. Because it is essentially logarithm taken of a concentration variable, one can directly include the changes in the pH value among the variables, $z_{\mathrm{pH}} = \Delta\mathrm{pH}$.

- **Dissipation.** It has been assumed that the systems being studied are in thermodynamic balance. This homeostasis can be extended, however: The steady state can be determined not only in terms of the variables, but also in terms of their derivatives. This means that one can study *dissipative systems,* where the rate of change remains constant, a constant flow of chemical flowing into or out from the system. Looking at the formula (1.3), it is clear that model linearity is not lost if one has variables like $z_{\dot{C}} = \Delta\dot{C}/\bar{\dot{C}}$.

- **Physical phenomena.** It is evident that structures that are originally linear, like phenomena that represent diffusion between compartments, etc., can directly be integrated in the model, assuming that appropriate variables (deviations from the nominal state) are included among the variables.

In strong liquids one cannot always apply concentrations, but one has to employ *activities* instead, or actual activation probabilities. If it is assumed that these activities are some power functions of the concentration so that $\mathcal{A} = a_1 C^{a_2}$, after taking logarithms the model still remains linear in terms of the original concentrations. This means that — even though linearity is not compromised — the variables may become multiplied by some unknown factors, so that there is some scaling effect.

The vector z selected here is the measurement vector, containing *all* possible quantities that can affect the system behavior — internal system variables and external environmental variables alike. To have the actual *data vector* to be employed in modeling, the vectors first have to be preprocessed and appropriately scaled — these issues are studied in chapter 2. In practice, specially if the relationships between units are not clear, it can be motivated to carry out explicit data normalization to make data items better compatible (this issue is studied closer in chapter 3).

1.4 Case 2: Gene expression

However, the above studies are not the whole story — *genes* are an integral part of the metabolic system. Nature's way of implementing the genetic descriptions are mindbogglingly sophisticated — but when trying to capture the essence in those processes, perhaps one does not need to exactly stick to the nonidealities in the implementations? Here the goal of the system is seen as more important: The complicated mechanisms are only needed to reach the consistent balance among the system components.

1.4.1 Process of overwhelming complexity

Above, the domain of chemical reactions was studied and a coherent modeling framework was proposed for capturing the relevant variables that characterize the process state. However, the cell differs from other reaction vessels of organic chemistry: The genetic system is essentially a part of the metabolic system,

controlling it. What can be said about it, how can this level be integrated in the system model?

The information required for building the proteins and controlling the cell metabolism is stored in deoxyribonucleic acid (DNA); the operational units in DNA are called genes. The proteins are synthesized in the process called gene expression: First the gene sequence is coded into messenger RNA in the transcription process, and, after being further modified and transferred from the nucleus into the cytoplasm, this code is compiled by ribosomes into proteins in the translation process. These proteins are either used as building blocks in the cell, or they act as enzymes, catalyzing other processes.

What makes this gene expression process specially complicated, is the fact that there exist feedback structures all the way along the process: First, the RNA molecules and proteins are "postprocessed" by various mechanisms controlled by some other genes and chemicals; second, some of the enzymes act as *transcription factors,* explicitly affecting the activity levels of other genes. In each case, there exist various complicated mechanisms (chromatin packing and dismounting, gene activation and inhibition, protein phosphorylation, myristylation, and glycosylation, etc.) how the interactions among the actors are implemented. It seems to be a hopeless task to accurately model the individual processes that are related in these processes (however, there exist various attempts to do that, for example [54]), and it seems that the system of interactions needs to be abstracted.

The gene interactions have been modeled applying abstract causality structures — an example of genetic networks is given, for instance, in [78]. Many different model structures have been proposed, for example *neural networks* [82]. However, there exist no unique gene regulation pathways, processes having very different time scales and relevances — simple projections onto a graph form can be misleading. Here, in the neocybernetic spirit, the pancausality idea is applied: In steady state, the transients have decayed, and all reaction chains have found their balance; the temporal sequences have changed to simultaneous patterns of effect flow, practically all genes participating in this equilibrium. How to characterize the internal tensions among genes?

1.4.2 "Cybernetizing" a genetic network

The genetic control system is much too complex to be modeled explicitly. The only possibility is to look at the genetic system directly from outside, studying the overall net effects. Again, when concentrating on the final thermodynamic balance among tensions, the time-domain complexities can be circumvented. There exist some clues.

Stationarity and statistics

Abstract away individual actions and realizations of interactions in the network, and assume that the stationary state has been reached. Is there anything one can say about such a system in general terms?

It has been observed that there exist peculiar similarities among very different

kinds of complex systems. For example, it has been claimed (see [12]) and [5]) that distributions in self-organized complex networks statistically follow the *power law,* that is, there generally holds

$$z_j = cz_i^D \tag{1.9}$$

for some behaviors-related variables z_i and z_j, and constants c and D. Here, z_i stands for the free variable, and z_j is some emergent phenomenon related to the probability distribution of z_i. This power law dependency seems to govern all structures with fractal and self-organized structure For example, if z_i is the "ranking of an Internet page", and z_j represents "number of visits per time instant", the dependency between these variables follows the power law: There are some very popular pages, whereas there are huge numbers of seldom visited pages. As compared to Gaussian distribution, the power law distribution has "long tails"; the distribution does not decay so fast.

It is interesting to note that the power law distribution is closely related to another modern concept, namely *fractal dimension.* Assuming that the variable z_i represents some kind of "yardstick", determining the scale factor, and z_j represents the level of *self-similarity,* so that when one zooms the original pattern by the factor of $1/x_i$, there exist z_j copies of the original pattern (and this zooming process can be repeated infinitely), the fractal dimension of that pattern can be defined as

$$D = \frac{\log z_j}{\log z_i}. \tag{1.10}$$

When the pattern is simple, this definition coincides with the traditional ideas concerning dimension, but for complex patterns, non-integer dimensions can exist. Now, it is easy to see that, after taking logarithms, the parameter D in (1.9) closely corresponds to the fractal dimension for the networked system.

Multivariate nature and linearity pursuit

In the multivariate spirit, one can extend the single-variable formula (1.9) by including more variables — assume there exist μ of them:

$$1 = c' z_1^{D_1} \cdot \cdots \cdot z_\mu^{D_\mu}. \tag{1.11}$$

If there is only one variable z_i changing at a time, and one solves for z_j, this formula corresponds to (1.9). Furthermore, there can exist various such dependency structures — assume there are ν of them:

$$\begin{cases} 1 &= c_1 z_1^{D_{11}} \cdot \cdots \cdot z_\mu^{D_{1\mu}} \\ &\vdots \\ 1 &= c_\nu z_1^{D_{\nu 1}} \cdot \cdots \cdot z_\mu^{D_{\nu\mu}}. \end{cases} \tag{1.12}$$

Now, if one takes logarithm on both sides of the formula, one has

$$\begin{cases} 0 &= \log c_1 + D_{11} \log z_1 + \cdots + D_{1\mu} \log z_\mu \\ &\vdots \\ 0 &= \log c_\nu + D_{\nu 1} \log z_1 + \cdots + D_{\nu\mu} \log z_\mu. \end{cases} \tag{1.13}$$

It turns out that the multiplicative dependency has become globally linear — by only preprocessing the variables appropriately. To find a still simpler (locally applicable) model structure, one can further differentiate these equations around the nominal values $\bar{z}_i$, so that there holds

$$\left\{ \begin{array}{rcl} 0 & = & D_{11}\frac{\Delta z_1}{\bar{z}_1} + \cdots + D_{1\mu}\frac{\Delta z_\mu}{\bar{z}_\mu} \\ & \vdots & \\ 0 & = & D_{\nu 1}\frac{\Delta z_1}{\bar{z}_1} + \cdots + D_{\nu\mu}\frac{\Delta z_\mu}{\bar{z}_\mu}. \end{array} \right. \tag{1.14}$$

Again, the variables $\delta z_i = \Delta z_i/\bar{z}_i$ are the relative deviations from the nominal state. It turns out that there again holds the linear dependency, for variables having been preprocessed in an identical manner as in (1.7),

$$0 = \Gamma^T \, \delta z. \tag{1.15}$$

The scaling of the variables is not determined in a unique manner. As it turns out (in chapter 3) traditional normalization of the variable variances is motivated. Determining the "nominal state" is equally vague; if nothing else is known, it has to be assumed that extensive information of the system variables has been acquired, and the mean values characterize the nominal state. In what follows, the "Δ" symbols are dropped for brevity; however, it is still assumed that variations around the nominal state are small.

Combining the results of this section and the previous one, it can be claimed that *all phenomena that are relevant for characterizing the cellular state can be captured in a homogeneous linear framework* as shown in (1.15). It is interesting to study the properties of data properties that are dictated by the assumed structure.

1.5 Probability interpretations

When the above simple formulations were derived, individual phenomena were abstracted away. There are no more individual samples or time points visible, only their statistical long-term properties. Thus, it is interesting to briefly elaborate on probability distributions.

1.5.1 Fractality revisited

Study the outlook of the *multivariate fractal distribution.* Assume that one of the variables has been expressed in terms of the other variables

$$\log z_j = -\sum_{i \neq j} \frac{D_i}{D_j} \log z_i. \tag{1.16}$$

Variable $\log z_j$ is a weighted sum of assumedly large number of assumedly independent stochastic variables $\log x_i$. Because nothing more accurately about these variables is known, it can be assumed, according to the Central Limit

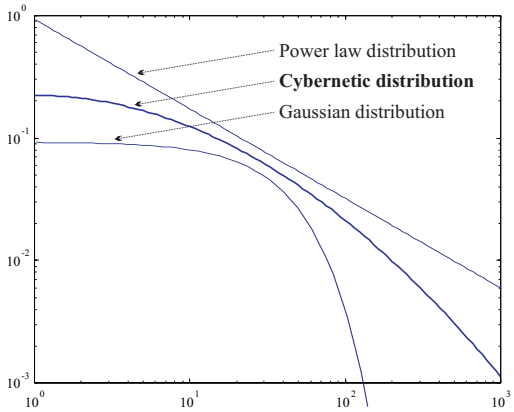

Figure 1.3: Schematic illustration of different distributions on the log/log scale

Theorem, that $\log y$ has normal distribution (or, indeed, *lognormal* distribution — see [55]):

$$p(\log z_j) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\left(\log z_j - \mu\right)^2 / 2\sigma^2\right). \tag{1.17}$$

Here, parameters c, μ, and σ are free parameters characterizing the outlook of the distribution. Taking logarithms, there holds

$$\log(p(\log z_j)) = c - \left(\log z_j - \mu\right)^2 / 2\sigma^2. \tag{1.18}$$

This means that the multivariate fractal distribution is *parabolic* rather than linear on the log/log axis, the three parameters being c, μ, and σ^2 (see Fig. 1.3).

Assume that the system variables can truly be characterized as having the dimension of *probability,* that is, genetic activity of a single gene can be seen as a probabilistic phenomenon. In such a case the above result gives new intuition. Indeed, the result is in conflict with "traditional" assumptions concerning fractal networks! This assumption seems to be supported also by evidence: For example, in Fig. 1.4, a manifestation of properties of a complex network are illustrated. It is clear that a parabolic curve better fits the observation points — the new model suits structures that are not strictly *scale-free.*

1.6 About more complicated distributions

In practice, model linearity cannot always be reached by as simple preprocessing of the variables as was presented above. For example, some genes can only be active in the vicinity of some location in the chemical data space — getting farther from that location in *any* direction makes the activity decay. The model structure can still be kept linear by appropriate selection and preprocessing of the variables; the key issue is to analyze how the system sees its environment.

If it is assumed that behaviors are results of high numbers of components interacting, the model is multiplicative with respect to the concentrations or probabilities. If logarithmic quantities are studied, one has additive models — one can assume that this assumption of log-linear behaviors can be extended beyond the

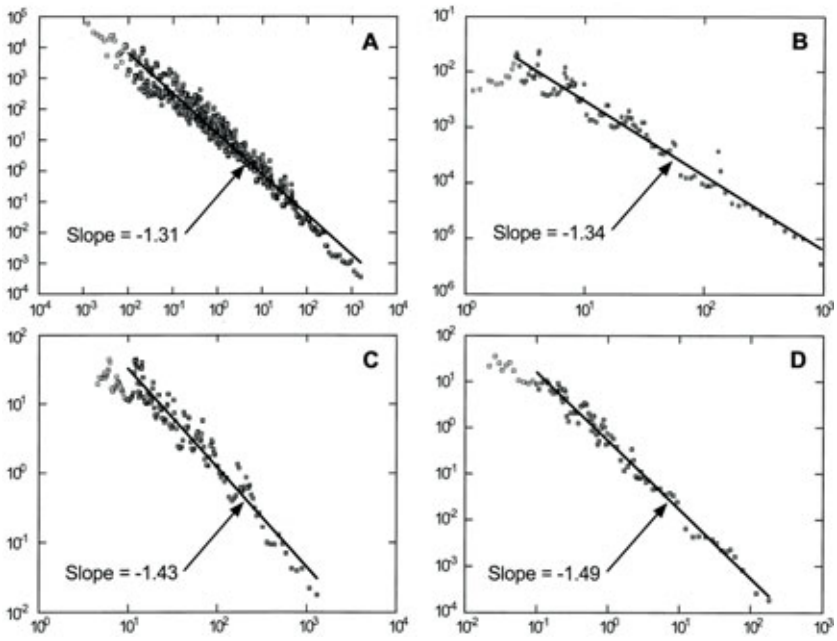

Figure 1.4: Properties of forest fires (from [52]): (A) 4284 fires on U.S. Fish and Wildlife Service lands (1986-1995), (B) 120 fires in the Western United States (1950-1960), (C) 164 fires in Alaskan boreal forests (1990-1991), and (D) 298 fires in Australia (1926-1991). The number of fires is given as a function of the burnt area

power law distributed variables. If the underlying distributions are Gaussian, one only needs to take into account the observations in the previous section: The behaviors can be assumed to be linearly related to *quadratic* functions of the input variables. This means that the set of input variables can (as the first approximation at least) be extended by including the squares of the most relevant variables, and products of them, among the input variables. Including the products of all variables among the features increases the size of the data space considerably.

This approach to representing general nonlinearities can be motivated in mathematical and in pragmatic terms. Mathematically speaking, smooth nonlinearities can be represented applying Taylor expansion, and such series can be approximated up to the second order when the quadratic terms are available. From the pragmatic point of view, the Gaussianity assumption is well in line with the Gaussian mixture model scheme (see chapter 6).

In this chapter, the homogeneity goal was reached: No matter what is the underlying realm like, the statistical properties of a complex network can be captured in a data vector, and it can be assumed that linear models are applicable if appropriate data preprocessing is applied. To truly capture all relevant variables, it is reasonable to include all variables that are potentially relevant — it is the task of the modeling machinery to select the most important of the variable candidates and to determine the dependency structures among them. Also, if

there exist nonidealities giving raise to further nonlinearities in the system, the data vector can be augmented by the appropriate feature variables hopefully capturing the structure of nonlinearities. This means that the data vector can become very high dimensional, and the models to be studied explicitly need to be robust against high dimensionality. This is a real challenge — specially as tackling among the multitude of variables should be done not manually but automatically.

The first step towards a model of complexity, and towards deeper understanding of biological systems was taken in this chapter. However, the model structure (1.7) is barren, being very descriptive, and it is not suitable for real applications. Next, in the Level 2, the same model is first extended to capture the cell-scale phenomena, and after that, a more suitable formulation, or the structure of the "emergent models" is presented.