# Model-Based Data Compression: From Data Compression to Information Condensation

Holger Wallmeier

Aventis Research & Technologies GmbH & Co KG, Core Technology Area Biomathematics,
Industrial Park Hoechst, G 515 A, D-65926 Frankfurt am Main, Germany
E-mail: wallmeier@CRT.hoechst.com

*Received: 23rd January 2001 / Published 11th May 2001*

## Abstract

Support of industrial research and development activities by computing and information technologies today is coupled to huge amounts of data. Therefore, data management is a very crucial aspect of successful application of information technologies. Various strategies are used to handle the situation, each of which has its merits depending on the type of data, the context, and the usage.

Apart from the very straightforward approach to distribute data on appropriate storage media of sufficient volume, there are three different 'philosophies' of data compression.

1. Non-lossy data compression
2. Lossy data compression
3. Model-based data compression

Types 1 and 2 are probably the most widely used because they do not necessarily introduce a bias into the compressed data. There are a number of methods known today that are fully reversible, or at least reversible to a large extent.

This is different for model-based data compression. The idea is useful for data being produced by dynamic, deterministic systems. Important is the existence of a model with well-defined data scheme and data structure. These model features can be used to condense the corresponding original data. Two examples from industrial research are presented.

First example is the representation of computer simulations of molecular ensembles by correlation functions. The second example is the representation of microbiological studies on pathogenicity by kinetic constants. In both cases, the underlying model together with methods to generate compressed data representations allows efficient interpretation of simulations or experiments, respectively.

High levels of data condensation provide a variety of opportunities to link results from research and development to auxiliary information from many different sources. Thus, powerful infrastructures for decision support can be created.

## Introduction

### Production of Data

The typical scenario of data generation starts from a device or automaton producing data (production), which will be recorded using some representation characteristic for the data and, of course, characteristic for the production itself. Based on this representation, data will be processed to extract the related information. In addition, the data may be transferred into a repository for later use.

Whenever the original production is deterministic, a faithful model of the original data production can be found, at least in principle. In such cases there is a unique data scheme and, furthermore, a well-defined analytical data representation. Usually, such models are given by a system of differential equations, the solutions of which define the data scheme. The way in which these solutions are determined also defines options of data representation. Extraction of information is then straightforward (Figure 1).

The advantage of the correspondence between original production and model production is that data scheme and data representation of the model

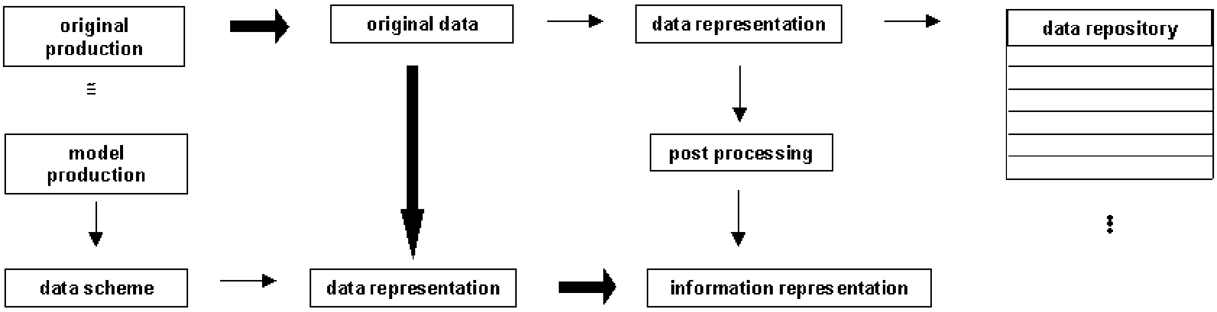


**Figure 1:** Data production and representation

can be used to generate a condensed representation of the original data. The information behind model data can usually be represented by few parameters.

The way in which this works will be shown by two examples. The first example is the computer simulation of molecular structures to analyze the stability of biomolecular complexes. In the second example, it will be shown how microbiological experiments with pathogenic bacteria can be analyzed in a very efficient way.

## Representation and Condensation of Data

Extracting and condensing information from data means creating a specific representation of the information. Basically, there are two different approaches to representing information. On the one hand, information can be mapped using predefined descriptor sets, thus creating specific profiles. On the other hand, information can be mapped in terms of relationships of the given object to known objects, which results, at least, in a delimiting view of the information. Genealogic aspects can be taken into account quite easily on a class and instance basis.

Both approaches offer several ways to condense information. Descriptor sets and profiles, for example can be handled using statistical methods such as clustering and classification, which also suggest strategies of visualization familiar from statistics and data mining.

Representing information by specifying relationships is first of all a simple and direct way of classifying objects based on similarity. In addition, this concept directly leads into the world of semantic networks.

## An Example from Molecular Modeling

Molecular modeling can be very useful to assess questions regarding, very generally speaking, stability and affinity of molecular systems. A very powerful, even though ‘expensive’ tool is the simulation of the dynamics of molecular systems. The underlying paradigm is based on perturbation theory. Simulations can be considered as computer experiments that allow the study of the response of a given system (molecular model) to some defined perturbation. The perturbation applied most frequently is just the kinetic energy of the $N$ particles associated with a given temperature $T$ according to [1]

$$E_{kin} = \frac{1}{2}\sum_{i}^{N} m_i \cdot \vec{v}_i^2 = \frac{3}{2}NkT \qquad (1)$$

Such simulations show the time evolution of the given system under the thermodynamic conditions specified and allow us to judge the stability of the given structural alignment, constitution, or conformation relative to some reference state. For this reason, simulation of molecular dynamics is a quite popular way of performing conformation-searches, especially for large molecular systems. By extending the analysis to the various aspects of entropy, affinity can also be estimated, at least on a molecular level.

## A Model for Dynamical Affinity of Molecular Systems

In practice, molecular dynamics simulations are performed by discretized integration of the respective equations of motion. [2] Since the characteristic frequencies of all relevant degrees of freedom must be resolved by the integration step-size, simulations of molecular dynamics are usually very lengthy and produce huge data sets (trajectories) if applied to large systems.

There is, however, a way to avoid very long simulations. The idea is based on the concept of collective modes of oscillation, which exist in stable molecular alignments. Indeed, the existence of such collective modes can be taken as a criterion for stability, because they make the difference between an unstable scattering state and a stable bound state of a molecular aggregate. According to quantum mechanics, their respective Eigenfrequency can identify such modes. Using a so-called Drude model, [3] which was originally developed for the electronic dispersion interaction of atoms and molecules by London, [4] this can be shown quite easily. Interacting molecules are represented by pairs of coupled harmonic oscillators (Figure 2). For simplicity, we take a pair of one-dimensional, coupled, identical harmonic oscillators positioned on the z-axis. The corresponding Hamiltonian is given by

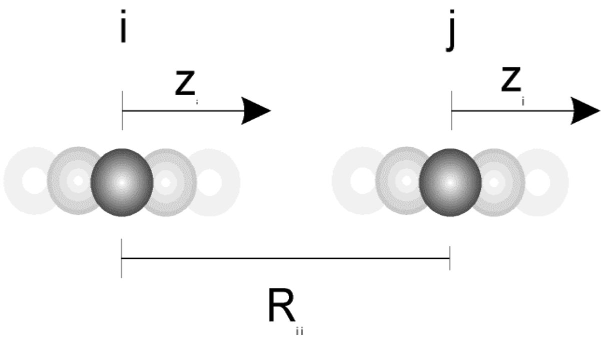


**Figure 2:** Coupled oscillators

$$H = T + V = \tfrac{1}{2}m\left[\dot z_i^2 + \dot z_j^2\right] + \tfrac{1}{4}K\left[z_i^2 + z_j^2 + 2a\cdot z_i\cdot z_j\right] \tag{2}$$

where $m$ is the mass of each oscillator, $K$ the force constant and $a$ the coupling constant, which is a function of the distance between the equilibrium positions of the oscillators. After separation of variables one has

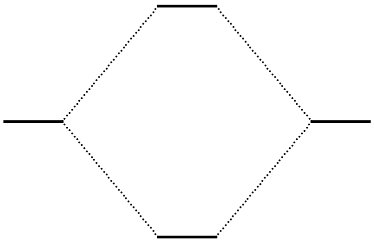
$$\begin{aligned} H &= \tfrac{1}{4}\left(\dot z_i + \dot z_j\right)^2 + \tfrac{1}{4}K(1+a)\left(z_i + z_j\right)^2 + \tfrac{1}{4}m\left(\dot z_i + \dot z_j\right)^2 \\ &+ \tfrac{1}{4}K(1-a)\left(z_i . z_j\right)^2 \\ &= H\left(z_i + z_j\right) + H\left(z_i - z_j\right) \end{aligned} \tag{3}$$

The first term represents the coherent motion of the center of gravity of the pair of oscillators and the second term the relative 'breathing' motion. Since both oscillators have a ground state frequency $\omega_0$, coupling results in a symmetric split of energy levels as shown in the following scheme (for $a > 0$)

$$\omega_+ = \omega_0 \cdot \sqrt{1+a}$$

$$\omega_0 = \sqrt{K/m}$$

$$\omega_- = \omega_0 \cdot \sqrt{1-a}$$



(for $a < 0$ in reversed order)

The energy of an oscillator is $\varepsilon = \frac{1}{2}\hbar \cdot \omega$, so that the splitting is given by

$$\varepsilon-\varepsilon_0=\tfrac{1}{2}\hbar\left(\omega_+ +\omega_- -2\omega_0\right) \qquad (4)$$
$$=\tfrac{1}{2}\hbar\omega_0\left[\sqrt{1-a}+\sqrt{1-a}\right]-2$$

For small $|a|$ one has

$$\varepsilon-\varepsilon_0\approx-\hbar\omega_0\cdot a^2 \qquad (5)$$

which is a typical second order, resonance-like effect.

Coming back to classical mechanics, one can calculate the sum over states ($|a| \ll \omega_0$) of the system

$$Q=\frac{kT}{\hbar\omega_+}\cdot\frac{kT}{\hbar\omega_-}=\frac{\left(\dfrac{kT}{\hbar\omega_0}\right)^2}{\sqrt{1-a^2}} \qquad (6)$$

The Helmholtz free energy is

$$A=-kT\cdot\ln Q=A_0+\frac{kT}{2}\ln\left(1-a^2\right) \qquad (7)$$

and since

$$S=-\left(\frac{dA}{dT}\right)_{\mathrm{V}}=k\cdot\ln Q+kT\cdot\left(\frac{d\ln Q}{dT}\right)_{\mathrm{V}},$$

and

$$S=S_0-\tfrac{1}{2}k\cdot\ln\left(1-a^2\right),$$

it is clear that

$$A-A_0=-T\cdot\left(S-S_0\right)=T\cdot\tfrac{k}{2}\ln\left(1-a^2\right)\approx\frac{kT}{2}a^2 \qquad (8)$$

Therefore, in terms of thermodynamics, coupling of oscillators adds a contribution to the energy of the overall system, which is mainly an entropy effect. It should be noted that the difference of the energy levels is independent of the sign of the coupling constant, since it is proportional to $a^2$. The energetic order of $\omega_+$ and $\omega_-$, however, is a function of the sign of the coupling constant. By analogy to the analysis of the (electronic) dispersion interaction by London, [4] this contribution to the entropy of molecular complexes can be called mechanical dispersion or, because of its stabilizing effect, dynamical affinity.

## Tracing Dynamical Affinity in Molecular Dynamics Simulations

In a molecular dynamics simulation dynamical affinity can be traced, mapping coherent and breathing motion by correlation functions.

$$G_{ij}=\frac{1}{T}\int_{t_0}^{T} g_i(t+\delta)\cdot g_j(t)\ ;$$

$$g=\begin{cases} & \textit{: position correlation function}\\ & \textit{: velocity correlation function}\end{cases}$$

*i, j : centers of correlation*

*$\delta$: correlation time*
*T-$t_0$: time of measurement (simulation time)*
*i = j : autocorrelation*
*i ≠ j : cross-correlation*
*$G_{ij}(0)$ = 1 : normalization*

For harmonic oscillators, one can define the autocorrelation functions for coherent and breathing motion

$$G_{ij}^{+}(\delta)=\frac{1}{T}\int_{t_0}^{T}\tfrac{1}{2}\left[g_i(t+\delta)+g_j(t+\delta)\right]\cdot$$
$$\tfrac{1}{2}\left[g_i(t)+g_j(t)\right]\cdot dt \qquad (9)$$

$$G_{ij}^{-}(\delta)=\frac{1}{T}\int_{t_0}^{T}\tfrac{1}{2}\left[g_i(t+\delta)-g_j(t+\delta)\right]\cdot$$
$$\tfrac{1}{2}\left[g_i(t)-g_j(t)\right]\cdot dt \qquad (10)$$

Now, it can be shown that the second derivative of these correlation functions is $-\omega^2$ for zero correlation time ($\delta$ =0). This means that the whole simulation can be condensed to just two independent numbers, $\omega_+$, $\omega_-$, and perhaps $\Delta\omega = \omega_+ - \omega_-$.

$G^+$ and $G^-$ are determined by selecting two centers (atoms or groups of atoms) i and j. The only condition to meet, is that i and j should be

| Molecule | $\Delta G$ kcal mol$^{-1}$ | $\Delta H$ kcal mol$^{-1}$ | $T\Delta S$ (297 K) kcal mol$^{-1}$ | $K_{binding}$ (M$^{-1}$) | $\Delta G_{Biotin}$ / $\Delta G_{HABA}$ | $T\Delta S_{Biotin}$ / $T\Delta S_{HABA}$ |
|---|---|---|---|---|---|---|
| Biotin | -18,3 | -32,0 | -13,7 | 2,5x10$^{13}$ | 3,5 | -2,0 |
| HABA | -5,27 | 1,70 | 6,97 | 10$^{4}$ | 1 | 1 |

**Table 1:** Thermochemical data of Streptavidin complexes with Biotin and HABA.

influenced in their dynamics by both, the coherent and the breathing mode.

Since $\omega_+$ and $\omega_-$ are determined from the trajectories of the full simulation ensemble, they are frequencies from the phonon spectrum of the whole system and not just frequencies of local molecular vibrations. In fact, splitting into $\omega_+$ and $\omega_-$ is a sensitive indication of the existence of a common, non-local mode of vibration for both oscillators. This of course shows that the interaction between the molecules has lead to a stable bound state and not an unstable scattering state.

## Streptavidin and Biotin

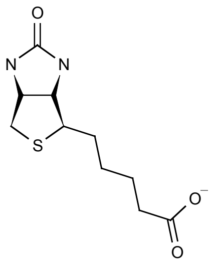
The example given below is a complex of two biomolecules, the protein Streptavidin and the vitamin Biotin. They form a specific complex with the largest binding constant known between biomolecules in nature. Therefore, this system is frequently used for immobilization of biomolecules. Surprisingly, experimental studies with molecules slightly different from Biotin show significant loss of stability and document the high specificity of the Biotin/Streptavidin complex.

For example, 2-(4’-hydroxyphenylazo)-benzoic acid (HABA) also binds to streptavidin, but with a binding constant which is 9 (!) orders of magnitude lower.

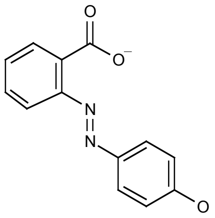
The thermochemical data measured for these complexes are given in Table 1. [5]

Apart from the remarkable values of the binding constants, it should be noted that the sign of the entropy contribution to the free binding energy changes going from Biotin to HABA. This is an indication of a change in the role of entropy.

Biotin



2-(4'-hydroxyphenylazo)-benzoic acid



$$K_{binding}^{Biotin} = \frac{[Streptavidin{:}Biotin]}{[Streptavidin][Biotin]} \cong 10^{13}\, M^{-1}$$

$$K_{binding}^{HABA} = \frac{[Streptavidin{:}HABA]}{[Streptavidin][HABA]} \cong 10^{4}\, M^{-1}$$

From the theoretical point of view, it is of course a challenge to model such a system. Fortunately, crystal structures of both complexes have been published. [6] Molecular dynamics simulations starting from these crystal structures have been run using the AMBER 3.0 [7] force field in NVT ensembles with water and counterions at 300 K. The ensembles have been thermalized during 30 psec simulations. Subsequently, another 15 psec were used to sample the trajectories from which oscillator correlation functions have been estimated. Table 2 summarizes the results of the simulations. For Biotin, four different orientations of the ligand in the binding pocket of Streptavidin have been simulated, for two HABA (Table 3). Columns 2 and 3 show the frequencies of the coupled oscillator motions derived from the autocorrelation functions $G^+$ and $G^-$. For the crystal structure orientation of Biotin (1) the coherent motion has the lower frequency and the breathing motion is significantly faster. In the first row of Table 2 the Biotin-results of a simulation without water and counterions are

| System / binding mode | $\omega_+$ (GHz) | $\omega_-$ (GHz) | $\Delta\omega$ (GHz) | Type of coupling | Splitting kcal.mol$^{-1}$ | T$\Delta$S (297 K) [6] kcal.mol$^{-1}$ |
|---|---|---|---|---|---|---|
| 1STP [8]/ 1[9] | 5.4 | 14.6 | -9.1 | a<0 | -0.87 | |
| 1STP / 1 | 2.9 | 12.4 | -9.6 | a<0 | -0.91 | -13.70 |
| 1STP / 2 | 6.1 | 0.76 | 5.4 | a>0 | 0.51 | |
| 1STP / 3 | 13.5 | 6.9 | 6.6 | a>0 | 0.63 | |
| 1STP / 4 | 10.7 | 2.2 | 8.4 | a>0 | 0.80 | |
| | | | | | | |
| 1HBA [8] / 1 | 8.7 | 4.7 | 4.0 | a>0 | 0.38 | 6.97 |
| 1HBA / 2 | 8.6 | 13.8 | -5.3 | a<0 | -0.50 | |

**Table 2:** Results of molecular dynamics simulations of Streptavidin complexes with Biotin and HABA. Starting from the crystal structures published, different orientations of the ligands have been studied. See text for further details.

given. The values do not differ very much from the results for the solvated system, which indicates the robustness of the method.

The interesting result is that the entropy contribution from oscillator coupling found in the molecular dynamics simulations shows the same relationship between Biotin and HABA as does the experimentally determined quantity *T·ΔS*. The agreement is 13% with respect to the experimental value, which is adequate for the force field chosen, the size of the simulation ensembles, and the simulation time.

$$\eta = \frac{T \cdot \Delta S_{Biotin}}{T \cdot \Delta S_{HABA}}$$

| Experiment | MD Simulation |
|---|---|
| 297 K | 300 K |
| -1.97 | -2.27 |

This underlines the role of oscillator coupling as indicator for stability of a given molecular alignment. At the same time it demonstrates the potential of data reduction that is given by this approach.

In terms of model-based data compression, we have the following situation. The original data production is the molecular dynamics algorithm in combination with the force field model of the system. The trajectories are the original data. Now, the model production is given by the coupled oscillators, the corresponding data scheme by the oscillator correlation functions, and the data representation by the oscillator frequencies. The representation of the information, *i.e.* the descriptor of stability, is given by the level splitting, calculated from the frequencies.

| Ligand orientation | Description | System |
|---|---|---|
| 1 | crystal structure | Biotin, HABA |
| 2 | upside down | Biotin, HABA |
| 3 | reversed | Biotin |
| 4 | upside down and reversed | Biotin |

**Table 3**: Orientations of the ligands Biotin and HABA bound to Streptavidin

## An Example from Biometry

Quite a different approach to model based data compression is possible in the area of kinetic studies of bacterial pathogenicity. Such studies are very important in infectious disease research. In a very general view, the key issue is the interaction between pathogens and the hosts they infect.

Besides the medicinal aspects of infection, pathogen-host interactions are the primary focus of target and lead compound search in the pharmaceutical industry. It is a complex phenomenon with several degrees of freedom.

## Dynamics of Infectious Disease Progression

Progression of an infectious disease is, in a generalized sense, always the result of several types of growth processes, which are characteristic for different phases of disease progression. [10] If one wants to identify targets for anti-infective drugs, the early phases of disease progression are of special interest.

The first phase is the invasion of the pathogen. This is some kind of transport phenomenon, which often is coupled to specific surface interactions and recognition steps.

What follows is a phase of establishment that usually results in a growth of the pathogen population. In this phase chemical communication between pathogen and host may occur, which can facilitate the pathogen's establishment significantly. The chemical 'messages' pathogens send to the host are called virulence factors. Typically, they serve to subvert normal functions of the host cells. Sometimes they have an immuno-suppressive effect. [11]

Next is the formation or enrichment of so-called pathogenicity factors. Very often they are toxins secreted by the pathogen. But also bacterial enzymes, which, for example cause necrotic degradation of host tissue belong to this class of factors.

Last but not least, the development of disease symptoms is related to the amount of pathogenicity factors formed. In all theses phases, however, there is some kind of host response to defend against the pathogen. For more complex host organisms it is an immune response.

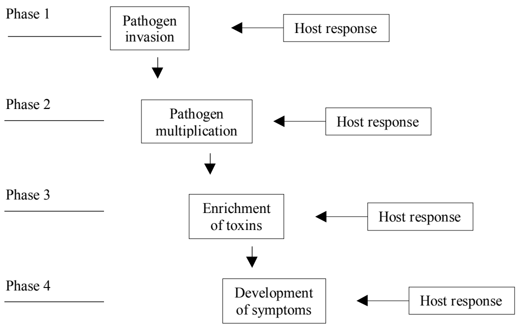


**Figure 3:** Phases of infectious disease progression. See text for details.

The scenario described above can be summarized in terms of the categories pathogenicity, virulence, and susceptibilty. Even though in literature pathogenicity and virulence are often used synonymously, a distinction based on genomical and disease progression considerations is possible. Pathogenicity is first of all a property of a pathogen that manifests in the formation of pathogenic factors like, for example toxins. [12] This, of course, depends on genotypic, as well as phenotypic conditioning of the pathogen. To be specific, what matters is type and amount of pathogenicity factors produced by the pathogen inside, or in contact with the host. The amount of factors formed, however, also depends on the size

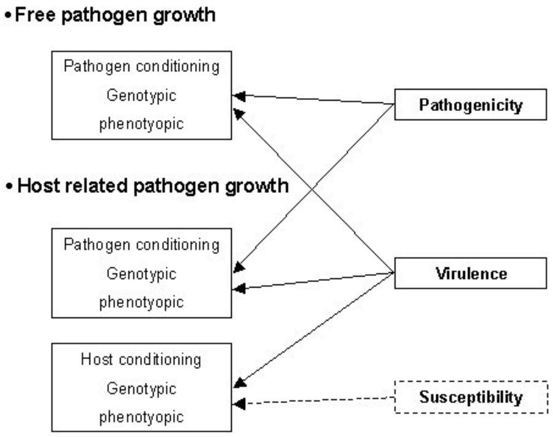


**Figure 4**: Pathogenicity, virulence, susceptibility, genotypic, and phenotypic conditioning.

of the pathogen population inside the host, which, in turn, depends on genotypic and phenotypic conditioning of the pathogen.

Due to host response, however, pathogen multiplication also depends on genotypic and phenotypic conditioning of the host. In principle, there are two degrees of freedom for the pathogen. These are, on the one hand its ability to produce pathogenicity factors, and on the other hand the size of population of pathogenicity factor producing pathogens inside the host.

Since virulence factors are often host specific, many authors refer the notion virulence to the combined effect of pathogenicity factor formation and population growth.

The third degree of freedom (see Figure 5) is the host's susceptibility to infection by the pathogen. Here, genotypic and phenotypic conditioning of the host are the important features.

Any research in the field of infectious diseases aimed at understanding the large variety of strategies pathogens have developed during evolution must analyze the kinetics related to the different phases. First of all, descriptors have to be identified that allow us to follow the individual processes by experimental measurements (see Table 4).

| Phase | Type of process | Descriptors |
|---|---|---|
| Invasion | - transport phenomenon, first/zeroth order kinetics;<br>- target recognition, signal transduction | invasive pathogen count, [13] optical densities of culture media<br>specific interactions |
| Pathogen multiplication | - free pathogen population growth<br>- invaded pathogen population growth (dependent on host response) | pathogen count, [13] optical densities<br>pathogen count, [13] disease marker concentration, antibody titer |
| Toxin enrichment | - secretion of toxins and other pathogenicity factors<br>- pathogen population growth | toxin concentration, antibody titer, disease marker concentration<br>pathogen count [13] |
| Development of symptoms | - host population decrease | disease marker concentration, antibody titer |

**Table 4:** Processes in disease progression

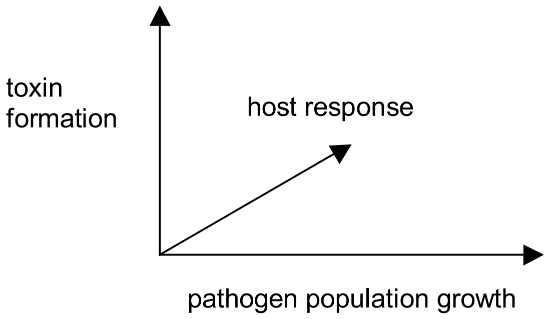


**Figure 5:** Genetic degrees of freedom in infectious diseases

A key problem in handling living organisms is reproducibility. Usually, this is taken care of by running replicate experiments and forming averages. In addition, time-resolved measurements are necessary to analyze the associated kinetics. To do so, the following model assumptions are useful.

## A Model for Infectious Disease Dynamics

The normal way to measure pathogenicity starts from a set of $N_0$ host organisms, which are infected. In the course of the experiment, decrease of the host population is measured. Typically, one obtains a sigmoid curve (Figure 6), which can be represented by the solutions of the following differential equation (DE). It is called the logistic, autocatalytic, or autokatakinetic differential equation [14]

$$\frac{dN(t)}{dt} = \left[k - g \cdot N(t)\right] \cdot N(t) \qquad (11)$$

describing growth processes with feedback. It is the equation of an exponential growth, which is modified by the second term in the square brackets. This second term depends on the population $N$ at time $t$ and constitutes the feedback. It can be agonistic ($g<0$), as well as antagonistic ($g>0$). The

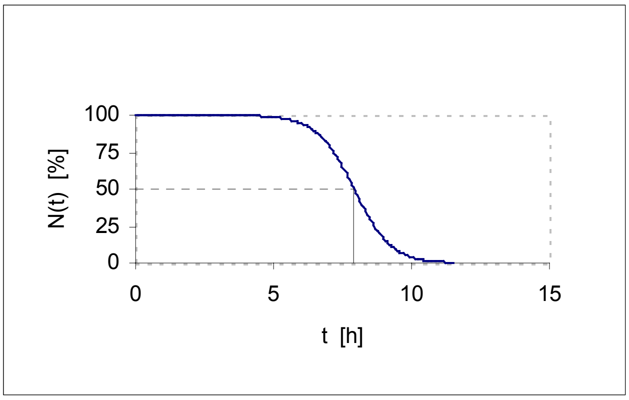


**Figure 6:** Decrease of a host population after infection. The time of the population's half-life is indicated.

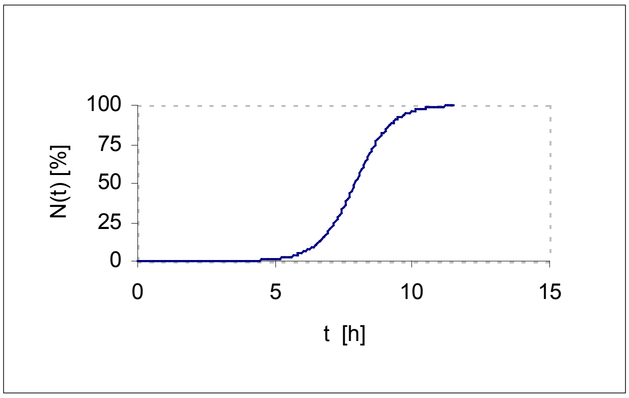


**Figure 7:** Increase of a pathogen population after infection.

general form of the solution is

$$N(t) = \frac{k \cdot N_0 \cdot e^{k \cdot t}}{k + g \cdot N_0 \cdot \left(e^{k \cdot t} - 1\right)} \qquad (12)$$

With the integration constant $N_0$, the initial size of the population, plus the rate constant *k,* and the feedback constant *g* there are three independent parameters. The combination of $N_0>0$ and a negative value of *k* describes the decrease of a population (Figure 6).

In contrast, vanishing $N_0$ together with a positive value of *k* describes a population that grows into a saturation state. With such parameters a growing pathogen population can be described (Figure 7).

Applying this equation to infection experiments, one has, first of all, equation (11) for the decrease of the host population. According to the considerations outlined above, it is easy to imagine that the kinetic constant *k* in fact depends on the growing pathogen population *P(t)* and is thus a pseudo-constant. Therefore,

$$k = k\left[P(t)\right] \qquad (13)$$

The growth of the pathogen population may either be unrestricted (free exponential growth)

$$\frac{dP(t)}{dT} = \kappa \cdot P(t), \text{where } P(t) = P_0 \cdot e^{\kappa \cdot t} \qquad (14)$$

or restricted,

$$\frac{dP(t)}{dT} = \left[\kappa - \lambda \cdot P(t)\right] \cdot P(t), \quad \text{where } P(t) = \frac{\kappa \cdot P_0 \cdot e^{\kappa \cdot t}}{\kappa + \lambda \cdot P_0 (e^{\kappa \cdot t} - 1)} \qquad (15)$$

reaching a saturation level due to host response. As

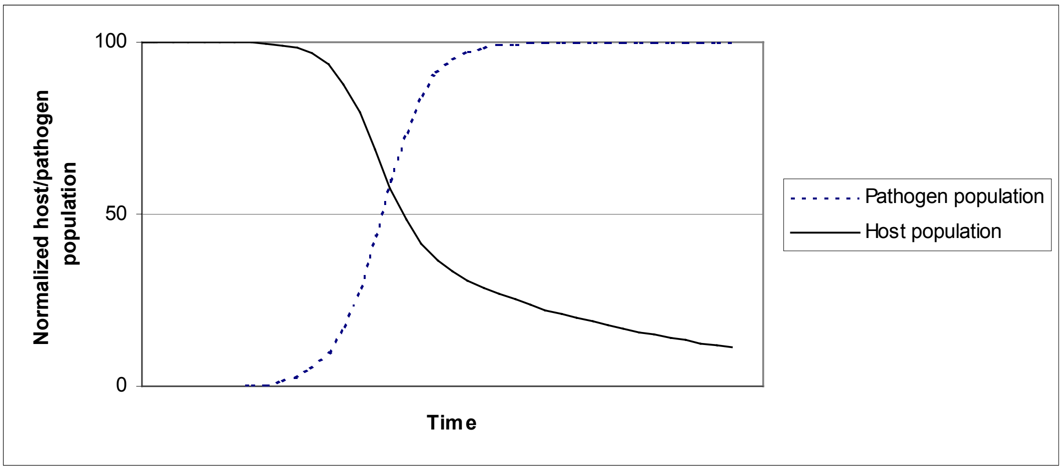


**Figure 8:** Decrease of host population coupled to growing pathogen population.

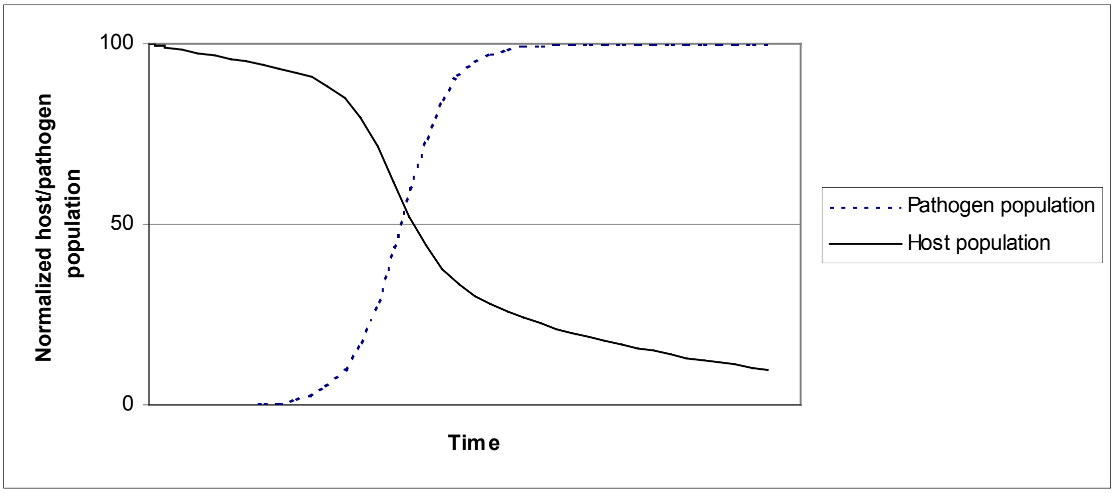


**Figure 9:** Decrease of host population coupled to growing pathogen population. A clear modulation of the host curve by the feed-back term can be seen

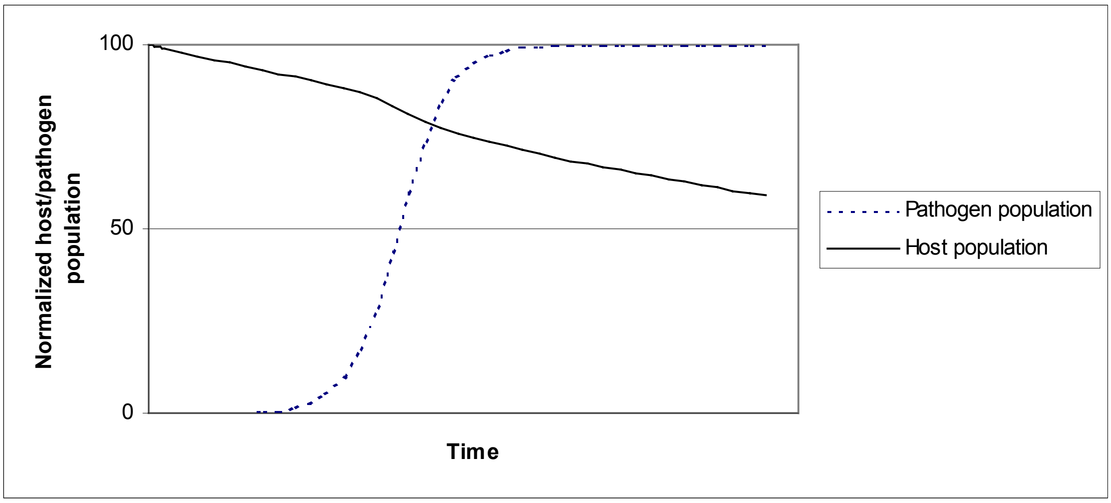


**Figure 10:** Decrease of host population coupled to growing pathogen population. Feed-back overrides the effect of the pathogen population

mentioned above, $P_0$ is small, and $\kappa$ is positive.

The simplest form of combining the two processes is to set

$$k = -\eta P(t), \qquad (16)$$

which, for example results in the situation shown in Figure 8.

It is obvious that the effect of the growing (not constant) pathogen population can be seen as a deformation of the host population curve. The degree of deformation increases with $\eta$. There is, however, a further type of deformation of the host population curve. It comes from the feedback term and can be seen in Figures 9 and 10. This certainly reflects host conditioning.

## Practical Applications

In experiments with time-resolved measurements, one usually has data reflecting the decrease of a host population that consists of test organisms such as, for example insects, mice, rats, or nematodes [15] (Figure 11).

Traditionally, the simplest way to measure pathogenicity is to count the host population after some predefined time $t_{scoring}$, which gives an *ad hoc* score as a percentage.

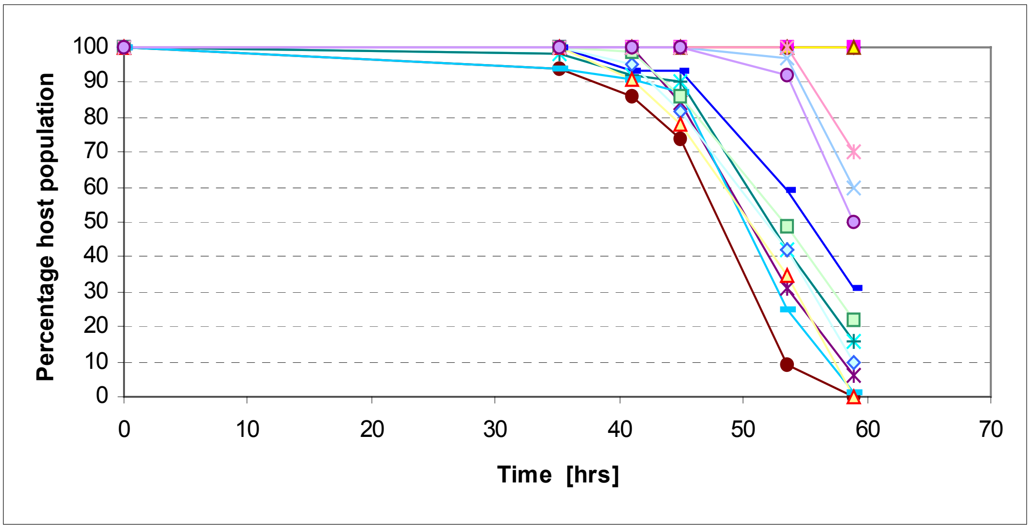


**Figure 11:** Time-resolved measurements of a C. elegans population infected by P. aeruginosa.

$$S_{adhoc} = \frac{N_0 - N_{t_{scoring}}}{N_0} \cdot 100\% \qquad (17)$$

Unfortunately, this way of measuring pathogenicity depends very much on the choice of $t_{scoring}$. Standardization is accomplished by normalization with the wild type of the pathogen:

$$S_{normalized} = \frac{\dfrac{N_0 - N_{t_{scoring}}}{N_0}}{\dfrac{N_0^{wildtype} - N_{t_{scoring}}^{wildtype}}{N_0^{wildtype}}} \cdot 100\%$$

(18)

Very often, time series are run until the host population has reached half its original size and

$$t_{1/2} = t\left(N_0/2\right) \qquad (19)$$

is taken as the measure of pathogenicity. This condenses the whole series of measurements to one single value. For a given host organism those pathogens with a low $t_{½}$ are more pathogenic than those with a higher value.

## Further Developments

In general, however, this is not sufficient to distinguish all possible effects that may modulate the interactions between a host and a pathogen. Coming back to the solutions of the logistic equation, steepness of the population curve at $t_{½}$ can tell a lot about pathogens, as well as hosts. [15] To improve the analysis, one has to fit solutions of the logistic equation (12) to the experimental data. This can be done, for example using the method by Marquardt and Levenberg. [16] The set of parameters $k$, $g$, and eventually $\kappa$; $\lambda$, or even $\eta$ allow the identification of those bacterial mutants that show extraordinary behavior. This allows scanning the genome for so-called pathogenicity and virulence genes. Furthermore, different mechanisms of infection can be distinguished.

Together with the huge amount of genomic bacterial information available in the near future, such methods can be used to look for entirely new ways of fighting infectious diseases. One can, for example try to target genes or gene products involved in the very first step of an infection. This would not kill a pathogen, but disable its establishment and later on multiplication in the host. Such a ‘gentle’ way of infectious disease prevention is very likely not to trigger the development of resistances. Since the pathogens can survive ‘outside’ the host, there is only little selective pressure.

Based on the situation described above, strategies for fighting infectious diseases must be defined. This also defines the type of targets to be searched and later on has an impact on the assays used for identification of active substances (lead compounds).

The normal strategy in target finding is to deactivate (knock out) genes systematically and to check by suitable assays with model organisms, to what extent pathogenicity, virulence, and, perhaps susceptibility are affected. Both steps are rather critical and need careful evaluation of the data generated and a very critical assessment of the results obtained.

## Summary

Whenever it is possible, model-based data compression serves two purposes. First of all it can be a great help to condense even huge data sets to very few numbers. Furthermore, the definition of the model necessary for compression is a very challenging step that often helps to gain deeper insights into the matter. It can reveal inconsistencies and facilitate the recognition of unknown phenomena. Together with the condensed data it offers possibilities to represent the information behind the data in a very efficient way. However, any kind of modeling is an abstraction and idealization. A model always has to skip part of the reality. This of course limits the applicability of model-based data compression and defines the due diligence that must be applied using it.

## References and Notes

[1] McQuarrie, D. A.;*Statistical Mechanics*, Harper & Row, New York, **1976**.

[2] van Gunsteren, W. F.; Weiner, P. K. (Eds.), *Computer Simulation of Biomolecular Systems*, ESCOM, Leiden, **1989**.

[3] Drude, P. K. L.; *The Theory of Optics*. Longman, London, **1933**.

[4] London, F. *Z. Physik* **1930**, 63, 245; *Z. phys. Chem. B* **1930**, 11, 222; *Trans. Faraday Soc.* **1937**, 33, 537.

[5] Weber, P. C.; Ohlendorf, D. H.; Wendoloski, J. J.; Salemme, F. R. *Science* **1992**, 243, 85.

[6] Weber, P. C.; Wendoloski, J. J.; Pantoliano, M. W.; Salemme, F. R. *J. Am. Chem. Soc.* **1992**, 114, 3197.

[7] Weiner, S. J.; Kollman, P.A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. *J. Amer. Chem. Soc.* **1984**, 106, 765.

[8] Brookhaven PDB entry code

[9] ‘in vacuo‘ (without water and counterions) simulation of the crystal structure of the Streptavidin/Biotin complex

[10] Finlay,B. B.; Falkow, S. *Microbiol. Rev.* **1989**, 53, 210.

[11] Mescas, J.; Strauss, E. J. *Emerg. Infect. Diseases* **1996**, 2, 271.

[12] Toxin is considered here to be not only a substance produced endogenically by the pathogen, but also toxic substances that are formed by host response to the pathogen.

[13] Measured as CFUs (colony forming units). See, *e.g.* Brock, T. D. *Biology of Microorganisms*, Prentice Hall, London **2000**.

[14] Dost, F. H. *Grundlagen der Pharmakokinetik*, Georg Thieme Verlag, Stuttgart **1968**; Nisbet, R. M.; Gurney, W. S. C. *Modelling fluctuating populations*, Wiley, New York, **1982**; Skehan, P. *Growth* **1986**, 50, 496; Renshaw, E. *Modelling biological populations in space and time*, Cambridge University Press, Cambridge, **1991**; M. Marusic, M.; Bajzer, Z.; Vuk-Pavlovic, S.; Freyer, J. P. *Bull. Mathem. Biology* **1994**, 56, 617.

[15] Mahajan-Miklos, S.; Tan, M. W.; Rahme, L. G.; Ausubel, F. M. *Cell* **1999**, 96(1), 47; Mahajan-Miklos, S.; Rahme, L. G.; Ausubel, F. M. *Mol. Microbiol.* **2000**, 37(5), 981.

[16] Dennis, J. E.; Schnabel, R. B. *Numerical Methods for Unconstrained Optimization and Non-linear Equations*, Prentice-Hall, **1983**. Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes in C*, Cambridge University Press, Cambridge **1992**.