9
Quantitative Analysis of Biochemical Data

The results of biochemical investigations can only rarely be interpreted without some form of quantitative analysis of the experimental data. In this chapter, we describe methods that can be used for such analysis taking typical biochemical topics such as enzyme kinetics and the thermodynamics and kinetics of molecular interactions as our examples. The aim of the computer-based exercises in this chapter is to provide the reader with direct experience of methods of data analysis that, we hope, will enable them to apply these approaches to their own data. We also include a short revision of the essentials of thermodynamics and kinetics relevant to the applications discussed.

9.1
Introduction

9.1.1
General principles of quantitative data analysis

The questions that arise when data are being analysed quantitatively are essentially the following:

- how well does the model under consideration, which is usually proposed on the basis of previous experience, perform in explaining the experimental data, bearing in mind the accuracy of that data? Is the model satisfactory, or is it necessary to consider alternatives?
- what values of the parameters characterising the system (rate constants, binding constants etc.) are most consistent with the experimental data?
- how accurate are these parameters, and what are the limits of error?

There are several important criteria that all procedures for data analysis should satisfy, chiefly:

- that the experimenter should be able to see the results of the analysis graphically to check whether they are reasonable, and get a feel for the accuracy
- that however the results are manipulated, the orginal raw data should not be lost

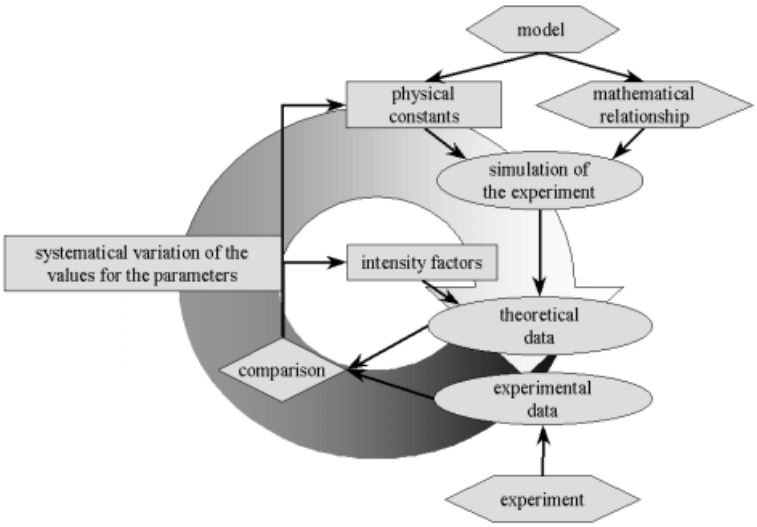


Figure 9-1. Schematic plan of the methods used in this chapter for quantitative data analysis.

- that there should be no hidden error propagation in the operations, for example by using transformations involving $1/x$, y^x and similar functions.

The basic concepts underlying the methods of data analysis discussed here are illustrated in Figure 9-1. The results of an experiment are data. A model is a description of the processes taking place in the experimental system being observed, which defines a mathematical relationship between the independent variables and the results. The model also defines physical parameters as variables to be fitted. With plausible initial values of the parameters, the mathematical relationships are used to obtain simulated data, which are compared with the experimental data. The values of the parameters are then varied until an optimal fit is obtained of the simulated and experimental results.

In the following sections, the basic concepts of quantitative data analysis are discussed, together with the terms used in the above scheme.

9.1.2 Experimental systems

The system is made up from various components and species. Components are molecules which differ in their covalent structures, e. g. enzyme, substrate and product; components can interact to form complexes, e. g. an enzyme-substrate complex. Species are all the entities present in the solution which differ in either their covalent or non-covalent structures; this will include components, complexes and, where relevant, different conformations of these.

9.1.3 Measurement and signals

In our analysis we consider a general relationship between the measurement or 'signal' and the composition of the solution; the signal is the experimental quantity being measured which gives information about the processes taking place. It is assumed that the total measured signal is additive in terms of the contributions of all of the species (i) in the solution, and furthermore, that the signal from each species is proportional to its concentration (c_i). Different species contribute to differing extents to the total signal, and the proportionality constant (f_i) is termed the intensity factor of the species i. This intensity factor defines the relationship between the concentration of the species and the measured signal (e.g. cpm, absorbance, fluorescence intensity, etc.). It is usual to have to take account of a non-specific but constant background signal, which we here define as the baseline (BL). The observed signal is thus given by the general Eqn. 9.1:

$$S = BL + \sum_i f_i c_i \tag{9.1}$$

So, for example, if we have the species A, B and AB in solution, then the signal is given by the expression:

$$S = BL + f_A c_A + f_B c_B + f_{AB} c_{AB} \tag{9.2}$$

To simplify the analysis, and to make the numerical analyses more stable, it is important that realistic assumptions are made about which species contribute to the signal, for example, in the case of radioactive detection, only those species that are labelled.

9.1.4 Models

A model represents an abstraction of the processes that are happening, or could be happening, during the experiment. From this model we can derive a mathematical relationship between the experimental results and the independent variables. Consider the simple case of the two species A and B forming a complex AB in a time-dependent process:

$$c_{AB} = f(t, c_A, c_B) \tag{9.3}$$

The experimental results would be the concentrations c_{AB}, and the independent variables would be c_A, c_B and the time t.

The model provides specific parameters for the fitting process, enabling theoretical data to be evaluated. These can be calculated either analytically or numerically. If the mathematical relationship between the signal (observation) and the parameters is sufficiently simple, it may be possible to obtain analytical solutions and calculate

the theoretical signal directly, i.e., in the present example to obtain values of c_{AB} knowing the initial concentrations of the concentrations $c^0_{A,}$ and c^0_B, and the time t.

However, in many cases, the mathematical relationships are not that simple, and analytical solutions may either not be possible in principle, or too difficult and cumbersome in practice. In such cases the theoretical data can be simulated by numerical methods.

A model is, of course, only a working hypothesis, whose validity is judged by its success in accounting for the data. If its performance is not satisfactory then alternative models should be sought or devised. However, if a model is to be replaced by a more complicated one, then it is important to check that the data really warrant this. More complicated models generally have more parameters, and more parameters will always lead to better fitting of the data. One should be guided here by the Principle of Parsimony, that other things being equal, the preferred model is the simplest one with the fewest parameters.

9.1.5 Selection of appropriate models

The choice of the right model to use to describe experimental results is one of the trickiest, and most interesting, tasks in scientific work, and this is a subject that can only be touched on here. As discussed above, we are guided by the Principle of Parsimony, that in science one should seek the simplest explanation for phenomena. In the present context, that means that we should define models with as few parameters as possible, consistent with obtaining a satisfactory description of the data. This is a sensible approach, because if a simple model fits the data adequately, then so necessarily must more complicated versions of that model. It follows that experimental observations can only serve to rule out models, often, but not always, because they are oversimplified; the data can never prove that a model is correct. The question naturally arises at this stage about how one can establish whether or not a model is successful in accounting for the data. There are several criteria for assessing the quality of a model.

- The absolute magnitude of the deviations between the theoretical and experimental data. Does the theoretical curve lie in the region of experimental uncertainty of the data points (taking particular care not to overestimate the accuracy of the data)?
- The direction of the deviations between the theoretical and experimental data. Are the deviations randomly distributed, sometimes above and sometimes below the curve, or are they clustered, above the curve in one region and below in another? If the deviations are not randomly distributed, this indicates that the theoretical curve is not a satisfactory fit to the experimental data. One reason for this is that the model is wrong and is not an adequate description of the situation; another is that systematic errors have been made in carrying out the experiment.

- Whether alternative models are available which can account for the data more satisfactorily.
- A good model should also have predictive power and suggest additional experiments which can be carried out to test the model further.

9.1.6 Parameters

Depending on the model under consideration, one obtains a set of parameters, that establish the relationship between the experimental data and the assumptions underlying the model. It is important to distinguish two kinds of parameter: global and local. This distinction is important when several data sets are being considered jointly in the analysis; the values of the global parameters must be the same in all cases, whereas those of the local parameters may vary from one data set to another.

- Global parameters: the values of the global parameters are the same for all of the data sets that are being considered in the analysis. We are dealing here with physical quantities such as binding or rate constants whose values we wish to determine.
- Local parameters: the values of the intensity factors discussed above can differ from experiment to experiment. Examples of intensity factors are: radioactivity ($CPM = f_i \cdot c_i$), fluorescence intensity ($signal = f_i \cdot c_i$), absorbance spectroscopy ($OD = f_i \cdot c_i$, in which f_i is the extinction coefficient of species i), ELISA ($signal = f_i \cdot c_i$) etc. Although the precise values of these factors, which are local parameters, are not particularly interesting in understanding the system, they are needed for the analysis.

9.1.7 Essential steps in the analysis

There are three basic steps in every data analysis (cf. Figure 9-1):

- arbitrary initial values of the parameters are introduced into the model to calculate theoretical concentrations for all of the species of interest in the system
- these theoretical concentrations are combined with initial values for the intensity factors to obtain theoretical values for the measurement or signal
- the values of the parameters and intensity factors are varied to obtain the best fit of the theoretical values of the signal to the experimental values; the combination of parameters which best fits the data is the result of the analysis.

9.1.8 Fitting data by the method of least squares

The classical method for fitting data to theoretical curves is linear regression. This procedure allows the equation of the best straight line fitting the experimental data to be calculated directly:

$$y = a + bx$$

$$\text{slope } b = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2} \tag{9.4}$$

$$y \text{ intercept } a = \frac{\left(\sum x^2\right)\left(\sum y\right) - \left(\sum x\right)\left(\sum xy\right)}{n\sum x^2 - \left(\sum x\right)^2}$$

Until relatively recently this was the only method that could be used conveniently to fit data by regression. This is the reason why so many classical approaches for evaluating biochemical data depended on linearising data, sometimes by quite complex transformations. The best known examples are the use of the Lineweaver-Burk transformation of the Michaelis-Menten model to derive enzyme kinetic data, and of the Scatchard plot to analyse ligand binding equilibria. These linearisation procedures are generally no longer recommended, or necessary.

In contrast to the explicit analytical solution of 'least-squares fit' used in linear regression, our present treatment of data analysis relies on an iterative optimization, which is a completely different approach: as a result of the operations discussed in the previous section, theoretical data are calculated, dependent on the model and choice of parameters, which can be compared with the experimental results. The deviation between theoretical and experimental data is usually expressed as the sum of the errors squared for all the data points, alternatively called the sum of squared deviations (SSD):

$$SSD = \sum_i \left(S_{i,\text{exp}} - S_{i,theo}\right)^2 \tag{9.5}$$

This deviation is now minimised by variation of the parameters. The combination of parameter values that 'best fit' the experimental data using this deviation as the criterion of best fit is the desired solution of the analysis. This process of finding a solution is termed 'iteration' because the solution is located by trying out many possible combinations of parameters; since the equations being fitted are in general non-linear, the process is more specifically one of iterative non-linear least-squares fitting.

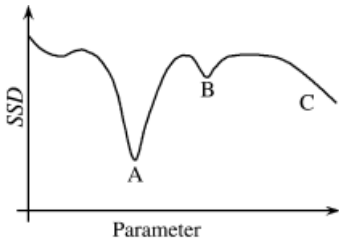


Figure 9-2. Two-dimensional representation of an error surface. Region A is the location of the global minimum, region B is a local minimum, and region C represents an area where the model is no longer valid and the slope of the error surface is directed away from the minimum.

The process is essentially as follows. All possible combinations of the parameters (physical constants and intensity factors), of number N, define an (*N+1*)-dimensional error space. Every point in that space has a characteristic value of the sum of the squared deviations (*SSD*), which thus generates an error surface in (*N+1*)-dimensional space. If, for simplicity, we consider a model with only two parameters, these can be represented on the *X* and *Y* axes, and the value of *SSD* on the *Z* axis. The error surface is now simply a surface in conventional three-dimensional space. An even simpler example with one parameter is illustrated in Figure 9-2 in which the parameter is shown on the *X* axis and the *SSD* is on the *Y* axis. The task in the fitting procedure is to locate the minimum value in the *SSD* curve (region A in Figure 9-2). It is impracticable to try out all possible values of the combined set of parameters, particularly when there are many of them. The procedure adopted in most computer programmes is, starting from initial values of the parameters (provided by the user) calculations are made of the slope (or derivative) of the error surface in (*N*+1)dimensional space. This is done by making a small variation of each of the parameters in turn and calculating the *SSD*. The programme then locates the region where the slope is steepest (downwards) and it alters the parameters by a small step in that direction to generate a new set of parameters, which fit the data better. From this new set of parameters, the programme repeats the operation in a second iterative cycle to locate the direction of steepest descent, and hence a new set of parameters.

This procedure depends on certain features that merit comment.

- The step length in the iteration is critical: if it is too short then the process of locating the minimum takes too long, whereas if the step length is too long the algorithm used in the programme can miss the target area, and thus never locate the minimum. The SOLVER algorithm used in Excel selects the step length automatically depending on the slope of the error surface and the result of the previous round of iteration.

- The 'result' located by the programme can be a local minimum (e.g. region B in Figure 9-2). Locating the global minimum is often not straightforward, particularly when the error surface is complex, and the programme can find itself trapped in a local minimum. The best means of avoiding this, or at least detecting when it is happening, is to begin the iteration process from different initial parameter estimates, and check whether the same solution is found in every case. If this does not happen, the solution with the lowest *SSD* corresponds to the best solution, although it should be noted that in some cases alternative solutions may be equally good in terms of their *SSD* values, bearing in mind the accuracy of the experimental data.
- To avoid local minima, most algorithms also test randomly selected points in the error surface. The extent to which a programme carries out these tests determines the speed of locating the minimum and the tendency of the algorithm to become trapped in local minima.
- All models have limits to their region of validity; for example negative values of rate or binding constants do not correspond to physically meaningful situations. In such regions, mathematical errors will arise, such as attempting to find the square root of a negative number, even though all of the equations have been correctly programmed.
- The slope of the error surface can lead the iterations into regions that are remote from the minimum. This situation can readily lead to failure to locate the minimum when the initial parameter estimates are not very good. To remedy this, a fresh set of initial estimates should be selected which fit the data better. In Figure 9-2, for example, it would be difficult to locate the minimum if the programme started in region C since the slope in the error surface is pointing in the wrong direction.

The usual criterion of 'best fit' is the sum of errors squared (the *SSD* discussed above) rather than the absolute magnitude of the errors. This procedure is mathematically justified when the errors in the data follow the Gaussian (or normal) distribution. Under these conditions the error distribution function is given by Eqn. 9.6 in which x is the measurement, μ the mean, and σ the standard deviation *cf.* Sect. 8.1.2:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{9.6}$$

When the data are distributed according to this function, the frequency of occurrence of data falls according to the square of the deviation. In practice, the sum of error square (*SSD*) criterion is also used in cases where it has not been explicitly established that the errors are normally distributed, and it appears to function quite well.

9.1.9 Global fitting of multiple data sets

If different sets of experiments have been carried out under circumstances where the observations depend on a common set of parameters, then it is sensible to attempt a global fitting of the data sets to obtain best estimates of the parameters. It is important here to distinguish clearly between the global and local parameters discussed above (Sect. 9.1.6). The global parameters are valid for all of the data sets and are fitted to all of the data, whereas the local parameters may assume different values for the various data sets. For example, if one were investigating a thermodynamic equilibrium, and monitoring the process using radioactive detection, the value of the equilibrium constant must be the same under the same conditions, whereas the specific activities of the reaction participants could well be different. In global data fitting, it is particularly important to keep the number of parameters as small as possible. There are two reasons for this. First, the general consideration that, following the Principle of Parsimony, one should seek to account for experimental data using the smallest number of variables. Secondly, that iterative fitting of the data becomes much more difficult (in fact, exponentially so) as the number of variables increases; the process becomes much slower, and there is an increasing risk that local minima will interfere with the fitting. To keep the number of parameters as small as possible, it is important to check, in particular, whether all of the local parameters are needed. For example, in the general case, it is assumed that all of the reaction participants contribute to the experimental signal or measurement (Eqn. 9.1), but if this is not in fact true, then it is better to set the intensity factors of as many species as possible to 0, and only allow the minimum number of species necessary to contribute to the signal. For example, in studies based on fluorescence detection, only species containing a fluorophore need to be assigned intensity factors.

One difficulty that can arise in global data analysis is that the signal intensities of different data sets can be very different. If the data are treated equally, this can lead to the situation that data sets or curves with high intensities completely dominate those with lower intensities, simply because their error squared parameters (SSD) are so much larger. The most effective way of dealing with this situation is to weight the SSDs of the different data sets or curves by a suitable factor, e.g. by the mean value of the data set, or the by the relevant intensity factor. It should be emphasised that weighting factors must never be treated as variables in the fitting process.

9.1.10 Introduction to error estimation

One of the most difficult tasks in day-to-day scientific activity is making reliable estimates of the errors and uncertainties in the data. How reliable are my data? How accurate are the parameters calculated from them? Can I, or should I, exclude particular models for explaining my data? These are examples of the sort of questions that need to be asked. We have already discussed the question of judging how well

models perform in accounting for data; we turn now to the question of assessing accuracy, on the basis that the model used is an appropriate one. It should be noted that we are dealing here with statistical errors and the treatment of outliers; systematic errors cannot be detected by these approaches (cf. Sect. 8.1.2). Three general strategies can be followed for error estimation.

- Statistical analysis of repeated measurements. If a very large number of data are available, then it is sensible to consider carrying out a rigorous statistical analysis. The simplest procedure is to do many replicates of the same experiment (or series of experiments, if more than one data set is needed for the analysis) and then analyse these independently. This is a very good way of assessing the error range (determined as standard deviations, maximal range etc.) of the individual parameters; the problem is the amount of work involved.
- Analysis of the accuracy of individual measurements. We are concerned here with the problem of assessing accuracy when the number of available data is limited. One means of gauging error is to remove individual data points from the fitting process to get a feel for the 'robustness' of the data. In effect, what this process does is to analyse the data on the assumption that the single experiment removed had not been carried out. It is possible in this way to assess how reliable the data are, and specifically to determine whether the outcome was highly dependent on the single result, implying that one would need to be very sure about it. This form of analysis is straightforward and revealing, and it ought to be a part of every data evaluation.
- Analysis of the shape of error surfaces. To conclude this section, we consider a more quantitative approach to error estimation. The first step is to estimate the accuracy of the individual data points; this can either be done by analysis of the variability of replicate measurements, or from the variation of the fitted result. From that, one can assess the shape of the error surface in the region of the minimum. The procedure is straightforward: the square root of the error, defined as the *SSD*, is taken as a measure of the quality of the fit. A maximum allowed error is defined which depends on the reliability of the individual points, for example, 30 % more than with the best fit, if the points are scattered by about 30 %. Then each variable (not the *SSD* as before) is minimised and also maximised. A further condition is imposed that the sum of errors squared (*SSD*) should not increase by more than the fraction defined above. This method allows good estimates to be made of the different accuracy of the component variables, and also enables accuracy to be estimated reliably even in complex analyses. Finally, it reveals whether parameters are correlated. This is an important matter since it happens often, and in some extreme cases where parameters are tightly correlated it leads to situations where individual constants are effectively not defined at all, merely their products or quotients. Correlations can also occur between global and local parameters.

9.1.11
Introduction to numerical integration

Kinetic processes can be described by differential equations; for example, for a reversible bimolecular association reaction:

$$A + B \underset{k_{21}}{\overset{k_{12}}{\rightleftharpoons}} AB \tag{9.7}$$

$$\frac{\partial c_{AB}}{\partial t} = k_{12} c_A c_B - k_{21} c_{AB} \tag{9.8}$$

This equation defines directly the change in concentration of the species AB with given concentrations of the reactants A and B, and the product AB. This is a differential equation whose solution is an expression of the form $c_{AB} = f(t, c_A^0, c_B^0)$. The solution involves a process of integration, which is often difficult, and sometimes impossible, at least analytically. In such cases, numerical integration can be used to simulate the time-dependent variation of c_{AB} in an experiment, enabling theoretical data to be obtained even for complex systems.

The procedure for numerical integration is as follows. Initial conditions are first selected: c_A^0, c_B^0, c_{AB}^0 and from this initial state the concentrations of the three component species are altered stepwise using 'fluxes' defined from the differential equation given above, with a finite time increment Δt.

Two different fluxes exist:

- F_{12}: 'association', $A+B \rightarrow AB$ for which $F_{12} = k_{12}\, c_A c_B \cdot \Delta t$
- F_{21}: 'dissociation', $AB \rightarrow A+B$ for which $F_{21} = k_{21} c_{AB} \cdot \Delta t$ (9.9)

The concentration changes are defined in terms of these fluxes as follows:

- $\Delta c_A = -F_{12} + F_{21}$
- $\Delta c_B = -F_{12} + F_{21}$ (9.10)
- $\Delta c_{AB} = -F_{21} + F_{12}$

from which new concentrations are obtained using the following general expression, in which $c_{i,old}$ is the 'old' concentration of the species (i) before the incremental change Δc_i :

$$c_i = c_{i,old} + \Delta c_i \tag{9.11}$$

The formulae given in Eqn. 9.9 are prototypes for bimolecular (F_{12}) and monomolecular (F_{21}) elementary reactions respectively. By combining these prototype equations, kinetic schemes of any desired complexity can be described and analysed.

9.2 Applications

9.2.1 Linear regression

Situations arise very often where data need to be fitted to linear equations. Linear regression is one of the classical procedures in general regression analysis, and before the advent of accessible non-linear fitting methods it was the only one that could be readily used. For n data pairs in the form (x,y) where y is a function of x, the linear equation of the form $y = a + bx$ that minimises the sum of errors squared (SSD) is given by:

$$\text{intercept } a = \frac{\left(\sum x^2\right)\left(\sum y\right) - \left(\sum x\right)\left(\sum xy\right)}{n\sum x^2 - \left(\sum x\right)^2}$$

$$\text{Slope} \quad b = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2} \tag{9.12}$$

(Exercise 2: Linear regression)

9.2.2 Michaelis–Menten kinetics

The Michaelis–Menten model shown below is the simplest mechanism for describing the kinetics of enzyme catalysed reactions:

$$S + E \underset{k_{21}}{\overset{k_{12}}{\rightleftharpoons}} ES \overset{k_{21}}{\rightarrow} E + P \tag{9.13}$$

According to this mechanism, the rate of the reaction depends on the rate constants k_{12}, k_{21}, and k_{cat}. In the simple mechanism shown above and with the assumption that ES is in a steady state K_M is defined as $K_M = (k_{cat} + k_{21})/k_{12}$. The dimensions of K_{cat} are concentration and (time)$^{-1}$ respectively. The rate of the reaction v (dimension: concentration/time) is given by the expression 9.14 and $v_{\max}$ is equal to $k_{cat} \cdot c_{E,total}$. The dependence of the reaction rate on substrate concentration is given by Eqn. 9.14, from which it can be seen that the k_M value is the concentration of substrate than gives half of the maximum rate $v_{\max} = k_{cat} \cdot c_{E,total}$. (*cf.* eq. 8.22)

$$v(c_S) = k_{cat} \cdot c_{E,total} \frac{c_S}{c_S + K_M} = v_{\max} \frac{c_S}{c_S + K_M} \tag{9.14}$$

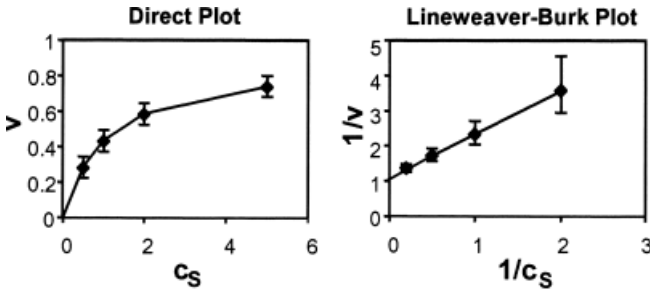


Figure 9-3. Error propagation in the direct analysis and Lineweaver–Burk analysis of Michaelis–Menten kinetics.

To evaluate K_M and k_{cat}, the rate of reaction is measured as a function of substrate concentration and the two kinetic parameters are determined using Eqn. 9.14. The classical method of doing this is by fitting the data to a linearised form of Eqn. 9.14 such as the Lineweaver-Burk plot shown in Eqn. 9.15 below: (*cf.* eq. 8.24)

$$\frac{1}{v(c_S)} = \frac{1}{v_{\max}} + \frac{K_M}{k_{cat}}\frac{1}{c_S} \tag{9.15}$$

From this it follows that a plot of $1/v$ against $1/c_S$ should give a straight line with an X-intercept of $-1/K_M$ and a *Y*-intercept of $1/k_{cat}$. The Lineweaver–Burk analysis illustrates very clearly the sort of problem that can arise when dealing with linearised data. An assumption that underlies simple linear regression following the procedure discussed in the previous section is that all of the data points have the same error, or specifically, standard deviation. This assumption is no longer valid when the data are transformed as is shown in the above diagrams. The diagram on the left illustrates a series of measurements where the data all have the same error; on the right, the same data are shown after transformation for Lineweaver–Burk analysis. It can be seen that the data points at low concentration (i.e. at high values of $1/v$ and $1/c_s$) have a much higher error than the other points, and the situation is made worse because these inaccurate points are also the ones that exert the most leverage on the linear regression, and hence on the derived kinetic parameters.

In the attached exercises we discuss three methods for analysing Michaelis–Menten kinetics:

- In the first approach, we examine the rate progress curves at various substrate concentrations, and use linear regression to evaluate initial rates. These initial rates are then fitted to the Michaelis–Menten equation (Eqn. 9.14) (Exercise 3: Michaelis–Menten kinetics I). This method has the advantage of being simple and robust. It has the disadvantage that the choice of data points used to obtain initial rates is often arbitrary, and also that the progress curves at low substrate concentrations show marked curvature because of substrate depletion.

- A further disadvantage of the above method is that linear regression is performed by the investigator and the least squares fit is carried out subsequently on the data derived from this linear regression. Thus the fit is not to the original raw data, and this is a situation that should be avoided if possible. This is not the case in the second approach to fit Michaelis–Menten kinetics (Excercise 7: Michaelis–Menten kinetics II). Here the Michaelis–Menten analysis is directly coupled to the linear regression and the fit is performed with the original data, thereby reducing the risk of operator subjectivity.
- The third approach uses an integrated form of Eqn. 9.14, which enables us to analyse the time dependence of product formation $c_P(t)$ directly to evaluate K_M and k_{cat} directly (Exercise 22: Analysis of Michaelis–Menten kinetics III). The integration is carried out numerically. This method allows data to be obtained from a single reaction progress curve, but it too suffers from some disadvantages, notably that many enzymes tend to lose activity in the course of an assay, and also that most enzymes show product inhibition. Both of these effects would cause pronounced curvature, reducing the rate of reaction and distorting the derived estimates of K_M and k_{cat}.

(Exercise 3: Michaelis–Menten kinetics I, Exercise 7: Michaelis–Menten kinetics II and Exercise 22: Michaelis-Menten kinetics III)

9.2.3 Dissociation kinetics

Dissociation reactions of the general form $AB \rightarrow A + B$ are monomolecular processes, in which the rate of decay of the complex is proportional to its concentration. The concentration dependence of c_{AB} is given by the following differential equation:

$$\frac{dc_{AB}}{dt} = -k_2, c_{AB} \tag{9.16}$$

which on integration yields Eqn 9.17 in which $c_{AB}(t)$ is the concentration of complex at any time t, and c^0_{AB} is the initial concentration at time $t = 0$:

$$c_{AB}(t) = c^0_{AB} e^{-k_{21} t} \tag{9.17}$$

Analysis of dissociation processes yields values for the rate constant k_{21}, whose dimensions are (time)$^{-1}$. This rate constant is related to the lifetime (τ) of the complex AB by the expression $\tau = (k_{21})^{-1}$, and to the half-life ($t_{1/2}$) by the expression $t_{1/2} = (\ln 2/k_{21})$.

(Exercise 4: Analysis of dissociation kinetics and Exercise 5: Global fitting of multiple data sets)

9.2.4 Binding data

The equilibrium constant for a simple bimolecular association process

$$A + B \overset{K_{Ass}}{\rightleftharpoons} AB \tag{9.18}$$

is defined by the expression:

$$K_{Ass} = \frac{c_{AB}}{c_A \cdot c_B} \tag{9.19}$$

This equilibrium constant is expressed as the association constant which has dimensions (concentration)$^{-1}$, in molar terms M^{-1}. The dissociation constant K_{Diss} is the reciprocal of K_{Ass} and has dimensions of concentration (M). The objective of the following derivation is to obtain an equation of the form $c_{AB} = f(c_{A,tot}, c_{B,tot}, K_{Ass})$, in which $c_{i,tot}$ are the total or stoichiometric concentrations of the components i (which are known), in contrast to the quantity c_i in Eqn. 9.19, which are the free concentrations of the species in solution, which are not known. An equation of this form will enable us to calculate theoretical data.

Using the conservation conditions: $c_{A,tot} = c_A + c_{AB}$ and $c_{B,tot} = c_B + c_{AB}$ Eqn. 9.19 can be written in the form:

$$K_{Ass} = \frac{c_{AB}}{(c_{A,tot} - c_{AB})(c_{B,tot} - c_{AB})} \tag{9.20}$$

The only unknown in this equation is the term c_{AB}. Expanding and rearranging Eqn. 9.20 yields the following quadratic equation:

$$c_{AB}^2 - c_{AB}\left(c_{A,tot} + c_{B,tot} + {}^{1}\!/\!_{K_{Ass}}\right) + c_{A,tot} \cdot c_{B,tot} = 0 \tag{9.21}$$

The solutions of a quadratic equation of the general form $x^2 + p\,x + q = 0$ are given by the two roots x_1 and x_2:

$$x_{1,2} = -\frac{\rho}{2} \pm \sqrt{\left(\frac{\rho}{2}\right)^2 - q} \tag{9.22}$$

In the present case only the negative square root term is physically meaningful, so the concentration of AB is given by the following equation:

$$c_{AB} = -\frac{c_{A,tot} + c_{B,tot} + {}^{1}\!/\!_{K_{Ass}}}{2} - \sqrt{\left(\frac{c_{A,tot} + c_{B,tot} + {}^{1}\!/\!_{K_{Ass}}}{2}\right)^2 - c_{A,tot} \cdot c_{B,tot}} \tag{9.23}$$

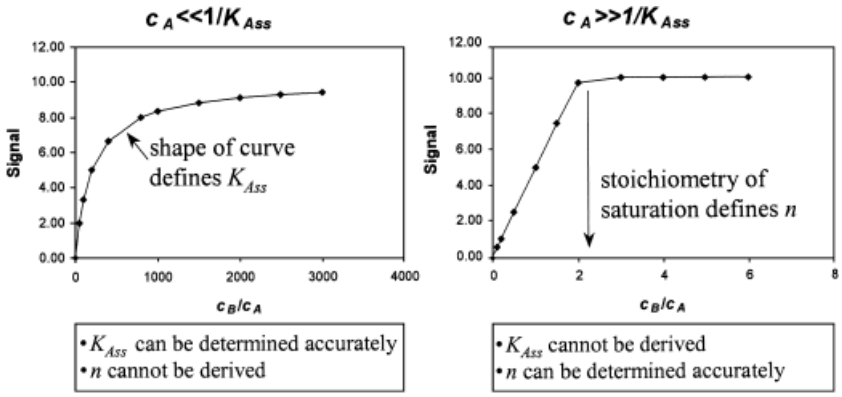


Figure 9-4. Typical results of a normal and stoichiometric titration binding analysis.

To determine values of K_{Ass} binding data are needed where the total concentrations of either A or B are comparable in magnitude to $1/K_{Ass}$.
(Exercise 8: Binding equilibria)

9.2.5
Independent identical binding sites

The above model and equations have to be modified if one of the species (say A) has several binding sites for the other species B. If the binding sites are independent and do not interact, then binding to each site on A can be described by Eqn. 9.19 given above. Taking all of the binding sites into account yields a hyberbolic binding curve whose binding equation only differs from Eqn. 9.23 in that for every molecule of A n binding sites exit such that the total concentration of binding sites is $n \cdot c_{A,tot}$:

$$c_{AB} = -\frac{n \cdot c_{A,tot} + c_{B,tot} + 1/K_{Ass}}{2} - \sqrt{\left(\frac{n \cdot c_{A,tot} + c_{B,tot} + 1/K_{Ass}}{2}\right)^2 - n \cdot c_{A,tot} \cdot c_{B,tot}} \tag{9.24}$$

To obtain accurate estimates of the number of binding sites (n), binding experiments (usually titrations) need to be performed under conditions where the total concentration of A is relatively high, specifically that $c_{A,tot} >> 1/K_{Ass}$; these conditions define a 'stoichiometric titration' where effectively all of the B added is bound until the sites on A are saturated. Titrations under these conditions are insensitive to the value of the association constant, so to obtain reliable estimates of K_{Ass}, data are needed from titrations at much lower concentrations, where $c_{A,tot} \leq 1/K_{Ass}$. It should be clear from this discussion that it is not easy to evaluate both n and K_{Ass} accurately, and it is usually necessary to do a global analysis of several data sets, obtained under different concentration conditions.
(Exercise 9: Independent identical binding sites I)

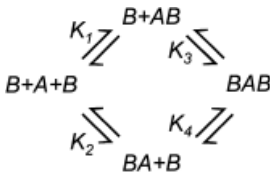
9.2.6
Analysis of simple binding data

The equation given above for n identical, non-interacting binding sites (Eqn. 9.24) is in principle soluble, although the solution is not straightforward. When binding is more complex and the sites are of different affinity and interacting, then analytical solutions cannot be obtained. However, analysis of the binding can be simplified by carrying out experiments under conditions where one of the interacting partners (say A) is present at a much lower concentration than the other. The concentration of the partner in excess (B) is varied, and the proportion of available binding sites on A which are occupied ($c_{AB}/c_{A,tot}$) is measured. The simplification in the analysis arises because the free concentration of B can be taken to be the same as the stoichiometric concentration (since $c_{AB} << c_B$). Eqn. 9.20 can be simplified considerably yielding, after inserting the conservation condition for A and rearrangement:

$$\frac{c_{AB}}{c_{A,tot}} = \frac{c_B K_{Ass}}{1 + c_B K_{Ass}} \tag{9.25}$$

9.2.7
Independent non-identical binding sites

Consider a macromolecule A that can bind several molecules of B. In the simplest case, where A possesses two binding sites for B, there are four possible species, A, AB, BA and BAB, whose concentrations depend on three binding constants:



(9.26)

Under conditions where $c_{A,tot} << c_{B,tot}$, terms involving the total concentration of A do not occur in the analysis (as shown above), and it is therefore not possible to use Eqn. 9.24 to analyse the stoichiometry of the binding equilibrium. However, even under these experimental conditions, it is possible to obtain information about the number of binding sites, provided the binding constants of the two processes are sufficiently different in magnitude.

(Exercise 10: Independent binding sites II)

Information about the minimum number of binding sites for B on the macromolecular species A can also be obtained if a signal can be measured which specifically monitors the concentration of A fully saturated with B (BAB in our scheme). For example, the enzyme DNA polymerase has two binding sites for metal ions, and both need to be occupied for the enzyme to be active. If it is assumed that the two

sites are independent, and hence $K_1 = K_4$ and $K_2 = K_3$ in Eqn. 9.26, the following expression can be derived for the occupancy of the two sites (designated 1 & 2):

$$\theta_1 = \frac{c_{AB}}{c_{a,tot}} = \frac{c_B K_1}{1 + c_B K_1}$$

$$\theta_2 = \frac{c_{BA}}{c_{A,tot}} = \frac{c_B K_2}{1 + c_B K_2} \tag{9.27}$$

and the proportion of A where both sites are occupied is given by:

$$\theta_{1,2} = \frac{c_{BAB}}{c_{A,tot}} = \theta_1 \cdot \theta_2 \tag{9.17}$$

In the special case of identical binding sites ($K_{Ass,1} = K_{Ass,2}$), the dependence of $\theta_{1,2}$ on the total concentration of A ($c_{A,tot}$) is weakly sigmoidal at low concentrations of B, and not hyperbolic; this is a direct indication that A can bind more than one B. The total concentration of bound ligand ($= \theta_1 + \theta_2$) follows a hyberbolic dependence, as expected since the sites are independent.

(Exercise 11: Independent binding sites III)

9.2.8 Cooperative binding

In the previous section, we discussed the case where the various binding sites were non-interacting; in this section we consider the other limiting case where A is either free, or fully occupied by B as the species AB_n, and the intermediate states AB, AB_2, ... , AB_{n-2} and AB_{n-1} are not populated. This behaviour arises because of positive interactions between the sites resulting in cooperative binding; according to Eqn. 9.26, cooperative binding occurs when $K_3 >> K_1$ and $K_4 >> K_2$. The model considered here represents 'all or none' behaviour, which is not just a theoretical model, but one which does actually occur with biopolymers.

In cooperative binding following the 'all or none' model

$$A + nB \overset{K_{Ass}^{app}}{\rightleftharpoons} AB_n \tag{9.29}$$

the association constant is defined by the expression:

$$K_{Ass} = \frac{c_{AB_n}}{c_A \left(c_B\right)^n} \tag{9.30}$$

Introducing the conservation condition for A with the further assumption that $c_{AB} << c_B$ yields the following equation:

$$\frac{c_{AB_n}}{c_{A,tot}} = \frac{\left(c_{B,tot}\right)^n \cdot K_{Ass}^{app}}{1 + \left(c_{B,tot}\right)^n \cdot K_{Ass}^{app}} \tag{9.31}$$

This equation describes a sigmoidal binding curve, where the degree of sigmoidal behaviour depends on the magnitude of n. The intrinsic binding constant of B for A (K_{Ass}) can be determined from the apparent binding constant (K_{Ass}^{app}) using the following relationship:

$$K_{Ass}^{app} = (K_{Ass})^n \tag{9.32}$$

(Exercise 12: Cooperative binding)

9.2.9 Association kinetics

The rate of a bimolecular association process $A+B \rightarrow AB$ is given by Eqn. 9.33:

$$\frac{dc_{AB}}{dt} = k_{12} \cdot c_A \cdot c_B \tag{9.33}$$

The rate constants for bimolecular association reactions have dimensions (concentration)$^{-1}$ (time)$^{-1}$. Although this differential equation has a very simple form, it does not have a very straightforward analytical solution. For this reason we use numerical integration methods to simulate theoretical data. This is a general approach that can be used to obtain solutions of complex kinetic processes. Although it is always easy to formulate differential equations like Eqn. 9.33, which express the time dependence of the various concentrations, solving the equations is another matter; it is often impossible to obtain explicit analytical solutions of the form $c=f(t)$ from which concentrations of the reaction participants can be directly determined. What can, however, be evaluated is the concentration change (or 'flux') for a species in a given time interval Δt under given conditions:

$$F_{12} = k_{12} \cdot c_A(t) \cdot c_B(t) \cdot \Delta t \tag{9.34}$$

The solution can be obtained by proceeding stepwise (using small values of Δt) and calculating $c_{AB}(t)$ using the expression:

$$c_{AB}(t + \Delta t) = c_{AB}(t) + F_{12} \tag{9.35}$$

This procedure is called numerical integration.
If we consider the following equilibrium:

$$E + S \underset{k_{21}}{\overset{k_{12}}{\rightleftharpoons}} ES \tag{9.36}$$

There are two different fluxes:

- F_{12}: $E + S \rightarrow ES$ for which $F_{12} = k_{12}\, c_E\, c_S \cdot \Delta t$
- F_{21}: $ES \rightarrow E + S$ for which $F_{21} = k_{21}\, c_{ES} \cdot \Delta t$ (9.37)

The concentration changes are defined as follows:

- $\Delta c_S = -F_{12} + F_{21}$
- $\Delta c_E = -F_{12} + F_{21}$ (9.38)
- $\Delta c_{ES} = -F_{21} + F_{12}$

from which new concentrations can be derived using the following expression in which, as before, $c_{i,old}$ is the old concentration of the species i before the new increment Δc_i:

$$c_i = c_{i,old} + \Delta c_i \tag{9.39}$$

(Exercise 17; Simulation of association kinetics using numerical integration, and Exercise 18: Analysis of association kinetics)

9.2.10 Pre-steady state kinetics

Enzyme reactions proceed, in general, via several intermediate states. A simple model incorporating multiple states is shown below: enzyme and substrate associate to form an enzyme-substrate complex, which undergoes a conformational change to $ES^{\#}$ before breaking down into enzyme and product.

$$E + S \underset{k_{21}}{\overset{k_{12}}{\rightleftharpoons}} ES \underset{k_{32}}{\overset{k_{23}}{\rightleftharpoons}} ES^{\#} \overset{k_{cat}}{\rightarrow} E + P \tag{9.40}$$

Since the concentrations of all the intermediate states are constant under steady state conditions, all of these states can, at least formally, be incorporated into a single kinetic intermediate state. It follows that under steady state conditions, kinetic data can provide no information about the existence and kinetic properties of intermediate enzyme-substrate complexes. An understanding of the mechanism of an enzyme catalysed reaction needs information about these intermediate states, which is therefore usually obtained from kinetic studies before steady state has been established, usually by rapid reaction methods. Comprehensive coverage of the techniques and methods of analysis of pre-steady state kinetics is beyond the scope of this chapter, but we discuss here methods for analysing simple exponential processes. Two approaches are used. In the first, the observed signal S(t) is fitted to an exponential function of the following form:

$$S(t) = Ae^{-\left(t/\tau\right)}\ ,\ \text{for decreasing signals} \tag{9.41}$$

$$S(t) = A\left(1 - e^{-\left(t/\tau\right)}\right),\ \text{for increasing signals}$$

A is the amplitude of the reaction and τ the time constant, with the dimension of (time). If the kinetic mechanism of the observed process is known then rate constants can be derived from the time constant. For example, for a simple dissociation process, such as the back reaction in Eqn. 9.36 but without the forward association process, the rate constant (k_{21}) is given by $1/\tau$. In this case, the value of τ is independent of reactant concentration.

If both forward and back reactions can take place, then $1/\tau$ depends on both k_{12} und k_{21}. In the special case that the concentration of S is much greater than that of E, then the association rate constant is given by the equation $1/\tau = k_{12}c_S + k_{21}$. Values of the two rate constants can be determined from the dependence of τ on the substrate concentration c_S; a linear regression of $1/\tau$ vs. c_S yields k_{12} as the slope of the plot and k_{21} as the Y intercept. For this analysis to be valid it is important to be sure that the observed reaction represents a single exponential process. If the reaction involves more than one exponential processes, then more complex models need to be considered, since the minimal number of reaction steps is given by the number of exponential processes.

(Exercise 14: Fitting rapid reaction data to exponential functions and Exercise 15: Error estimates for Exercise 14)

This method of analysis has several disadvantages, one of which is that intermediate parameters (τ) are evaluated from the data which then form the basis for global fitting of the data; consequently, the global fitting is not carried out on the raw data directly. A second drawback is that the predictive power of this analysis as regards mechanism is rather limited.

An alternative method is to use direct integration of the differential equations that describe the mechanism of the reaction. An advantage of this procedure is that the fitting is carried out directly to the raw data; a disadvantage is that numerical integration has to be used, since in most cases, particularly those of any kinetic complexity, the resulting systems of differential equations cannnot be integrated analytically.

(Exercise 20: Simulation of a complex enzyme catalysed reaction and Exercise 21: Analysis of the kinetics of a complex enzyme catalysed reaction)

9.2.11
pH dependence of enzyme catalysed reactions

The rate of an enzyme catalysed reaction is not only dependent on the concentrations of enzyme and substrate, but also on the conditions of the reaction. An important parameter affecting rate is the pH, defined as $-\log_{10}c(H^+)$, and it is very common that enzymes have a pH optimum. pH can have several effects: 1) protons may

participate in the catalytic reaction itself; 2) the protonation state of substrates and co-substrates may alter, with consequent effects on rate; 3) the protonation state of the enzyme itself may alter. In our example here, we deal with the last case. Proteins contain many groups that can undergo protonation-deprotonation reactions, including the N-terminal amino and C-terminal carboxyl groups and the side chains of the following amino acids: Asp, Glu, His, Cys, Tyr, Lys and Arg. The state of protonation of a group is conveniently represented by its pK_a value, which is the negative decadic logarithm of the dissociation constant for the protonation reaction:

- $K_a = \frac{c(H^+)c(A^-)}{c(HA)}$ (9.42)
- $pK_a = -\log_{10} K_a$

The proportions of a group A in the protonated $\theta(HA)$ and deprotonated $\theta(A^-)$ states can be evaluated using Eqn. 9.42:

$$\theta(A^-) = \frac{c(A^-)}{c(A_{tot})} = \frac{K_a}{c(H^+) + K_a} = \frac{10^{-pK_a}}{10^{-pH} + 10^{-pK_a}} \quad (9.43)$$

$$\theta(HA) = 1 - \theta(A^-)$$

The following questions are important for analysing the pH dependence of enzyme catalysed reactions:

- how many protonation reactions participate?
- which protonation state must the pH-sensitive groups on the enzyme be in?
- what are the pK_a values of these groups?

We consider a general model to analyse the protonation equilibria. If the enzyme possesses n groups that can participate in protonation-deprotonation equilibria, then in principle 2^n different species can be formed. For example, if $n = 3$, all three groups can be protonated (HHH), two (HH–, H–H and –HH), one (H–, –H– and –H) or none (—). The probability (P) of occurrence of these species, and hence their relative concentrations, depends on the product of the probabilities that each individual group is in a particular state:

- $P(HHH)$ = P(group 1 is protonated) × P(group 2 is protonated) × P(group 3 is protonated)
- $P(HH–)$ = P(group 1 is protonated) × P(group 2 is protonated) × P(group 3 is unprotonated)
- etc.

The probability of a group being in a particular protonation state is given by Eqn. 9.43, and combination of these probabilities multiplied by the total concentration of enzyme yields the concentrations of the different species.

The turnover rate is used as an 'effect' or signal to monitor the protonation, and thus the observed rates can be used to analyse the thermodynamic protonation equilibria. In the general case, every species would be assigned an intensity factor, and the signal (observed rate of the reaction) would be the sum of all of these factors. For our analysis, we make the simplifying assumption that only one species is catalytically active.

(Exercise 16: pH dependence of enzyme catalysed reactions)

9.2.12 Analysis of competition experiments

Competition experiments are widely used in the biosciences, particularly in studies of binding interactions. A simple example is shown in Eqn. 9.44.

$$AB + C \underset{K_{12}}{\rightleftharpoons} A + B + C \overset{K_{23}}{\rightleftharpoons} A + AC \tag{9.44}$$

In this example, the equilibrium between *A*, *B* and *AB* is affected by the addition of *C*. The popularity of the competition technique is due to the fact that it can be used to investigate interactions (in this case the binding of $A + C = AC$) without having to detect the participating species (free *C* and the complex *AC*). The method relies on using one interaction (here $A + B = AB$) as a reporter to monitor the other. This assumes, of course, that a suitable signal is available to follow the formation of *AB*. The diagrams below illustrate (left) the formation of *AB*, and (right) the effect of adding *C* to a system containing *A*, *B* and *AB*: on addition of *C* the species *AC* is formed at the expense of *AB* whose concentration falls, with a concomitant decrease in the observed signal.

It is also a desirable feature of competition experiments that they allow more precise comparison of the binding of different species (in this case *B* and *C*) to a common target (*A*) than is possible in separate binding experiments. It is also possible to use this approach with a single experimental set-up to test the binding of many different ligands to *A*, on the basis that these ligands all compete with *B* for the same binding site.

The analysis of coupled equilibria is the most complex problem that is considered in this chapter. It may seem surprising that such apparently straightforward systems like those shown in Eqns. 9.26 and 9.44 should present such great difficulties in analysis, the more so because it is a trivial matter to calculate the equilibrium constants, if the concentrations of the various species are known. However, that situation arises very rarely for several reasons:

- in most investigations only some of the species can be detected
- it is usually the case that only one 'signal' is measured, whose dependence on the concentration of reaction participants may be complex and must be derived from the model
- experimental error

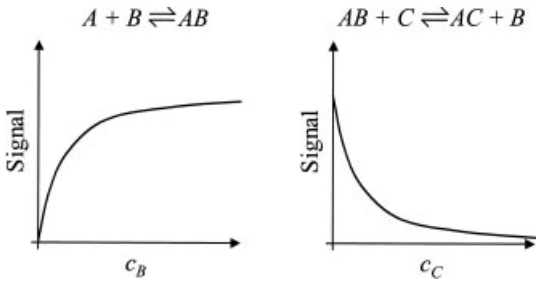


Figure 9-5. Indirect analysis of molecular interaction of *A* and *C* by competition.

Proceeding as we have done before with complex systems, we calculate theoretical data to deal with these systems. Since straightforward analytical solutions are not available, even for such simple cases as Eqn. 9.44, numerical methods are used to simulate solutions of the equilibria.

(Exercise 23: Analysis of competitive binding equilibria)

9.3 Guide to the CD

- The file 'chapter 9.pdf' contains the text of chapter 9.
- The file 'Introduction and theory.pdf' contains the Introduction.
- The file 'Guide to the exercises.pdf' contains advice about solutions to the exercises.
- The directory 'Solutions' contains programmed worksheets with solutions for all of the exercises.
- The file 'readme.xls' in the 'Solutions' directory describes the colour coding used in the cells in the accompanying Excel files.

Appendix I: SI-Units

Base units

physical quantity	name of unit	abbreviation
lenth	metre	m
mass	kilogramm	kg
time	second	s
current	ampere	A
temperature	kelvin	K
luminous intensity	candela	cd
amount of substance	mole	mol

Derived Units

physical quantity	definition	name of unit	abbreviation
area		square metre	m^2
volume		cubic metre	m^3
density	mass/volume		kg/m^3
specific volume	volume/mass		m^3/kg
molar mass	mass/amount of substance		kg/mol
concentration	amount of substance/volume		mol/m^3
molar concentration		molarity (M)	1M=1mol/l*
frequency	events/time	hertz	1Hz=1/s
force		newton	$1N=1kg\,m/s^2$
pressure	force/area	pascal	$1Pa=1N/m^2$
energy		joule	1J=1Nm
power	energy/time	watt	1W=1J/s
dynamic viscosity			Pa s
electric potential		volt	V
electric conductance		siemens	A/V
electric resistance		ohm	$1\Omega=1V/A$
electric charge	current · time	coulomb	1C=1A/s
electric capacity	charge/voltage	farad (F)	C/V
radioactivity	events/time	bequerel (Bq)	1/s
enzyme activity		katal (kat)	mol/s

* The litre is defined as $10^{-3}\ m^3 = 1\ dm^3$. This book uses the symbol l in preference to the alternative allowed symbol L.

Appendix II: Conversions into SI-Units

Force

	n (Newton)	dyne
1 N	1	$1 \cdot 10^5$
1 dyne	$1 \cdot 10^5$	1

Pressure

	Pa	bar	atm	Torr
1Pa (pascal)	1	10^{-5}	$0.986923 \cdot 10^{-5}$	$7.50062 \cdot 10^{-3}$
1bar (10^6 dyn · cm^{-2}	10^5	1	0.986923	$7.50062 \cdot 10^2$
1 atm	$1.01325 \cdot 10^5$	1.01325	1	760
1 Torr	$1.333224 \cdot 10^2$	$1.333224 \cdot 10^{-3}$	$1.315789 \cdot 10^{-3}$	1

Energy

	J	kWh	kcal	MeV
1J (joule)	1	$2.778 \cdot 10^{-7}$	$2.388 \cdot 10^{-4}$	$6.242 \cdot 10^{12}$
1 kWh (kilowatt hour)	$3.6 \cdot 10^6$	1	$8.598 \cdot 10^2$	$2.247 \cdot 10^{19}$
1 kcal (kilocalorie)	$4.187 \cdot 10^3$	$1.163 \cdot 10^{-3}$	1	$2.614 \cdot 10^{16}$
1 MeV (mega electron volt)	$1.602 \cdot 10^{-13}$	$4.450 \cdot 10^{-20}$	$3.826 \cdot 10^{-17}$	1