Chapter 3
Energetics and Dynamics of Biological Systems

While in the previous part of this book basic physical principles are explained governing the formation of molecular and supramolecular biological structures, we will come now to various functions of cells, tissues, organs, and organisms. For this, of course, molecular considerations form an important fundament, but at the same time, phenomenological parameters, like concentration, volume, viscosity, dielectric constants, conductivity, etc., are used which in fact are defined for large and homogeneous systems. In this way, we begin to enter the field of the so-called continuum physics.

At the same time, the approaches of quantum mechanics and statistical thermodynamics will now be replaced by those of phenomenological thermodynamics. These approaches also are defined for sufficiently large homogeneous phases. We should always be aware that in some cases, this in fact, does not meet the particular conditions of the biological system. We mentioned this problem in the previous part of this book, considering for example, mechanical or electrical properties of biological membranes. The step from a molecular to a phenomenological approach, nevertheless, is inevitable when considering larger systems like cells and organisms.

We will come back to this point in general in context with the electrical structure of organisms, discussing levels of biological organization in Sect. 3.5.1 (Fig. 3.33).

3.1 Some Fundamental Concepts of Thermodynamics

One of the first important treatise on the problems of thermodynamics was the famous monograph by Sadi Carnot, entitled "Réflexions sur la puissance motrice du feu et sur les machines propres a développer cette puissance" which was published in 1824. In this book, thermodynamics was introduced as the theory of heat engines. Julius Robert Mayer (1814–1878) was a physician and scientist to whom we owe

R. Glaser, *Biophysics*,
DOI 10.1007/978-3-642-25212-9_3, © Springer-Verlag Berlin Heidelberg 2012

the first numerically defined correlation between heat and work (1842). He therefore can be considered as the discoverer of the first law of thermodynamics. As a result of physiological observations he had already discussed how the functioning of the human body related to energy transformation in a heat engine. Meanwhile, thermodynamics has become the theoretical fundament of all kinds of energy transformation, and consequently, of all kinds of movement.

Applying thermodynamic considerations to open biological systems, however, requires an extension towards explanations of irreversible processes, i.e., towards nonequilibrium thermodynamics. This extension is made in two steps: Firstly, only small deviations away from the equilibrium are considered. In this case, linear relations between forces and rates can be assumed. In contrast to these *linear approaches*, the thermodynamics of *nonlinear* processes is applied to systems far from equilibrium. In this case, so-called *dissipative structures* appear, which are stationary states with completely new qualities.

It seems important to emphasize here that although the far-from-equilibrium condition of an organism represents an absolute precondition of life, nevertheless there exist many subsystems, which can be properly calculated using equilibrium thermodynamics, or thermodynamics of linear approaches. This means that biophysics must concern the whole scale of thermodynamic approaches.

3.1.1 Systems, Parameters, and State Functions

In Sect. 2.1.3, the term "system" was introduced in context with an explanation of the term "structure." We defined the system as an aggregate of elements with certain interrelations. Furthermore, we explained that systems, the interrelations of which are not simply relations, but interactions, are so-called dynamic systems. These are the kinds of system to which thermodynamic approaches are to be applied.

The question of what kind of model we should use, what we should consider as a system, and what are its elements, depends exclusively on the particular problem, and the corresponding point of view. An element of one kind of system can become a system in itself when calculating another problem. An organism, for example, as an element in an ecological system can become a system itself, if we ask a physiological question, such as for example the interactions between its organs. The organ can be considered as a system of cells, the cell as a system of organelles, and so on.

A dynamic system can be analyzed in different ways. In contrast to system theory, which calculates the kinetic interplay of individual elements (see the Introduction to Sect. 5), thermodynamics considers a system simply as a continuum which stands in a defined interrelation with its environment. The limit of this continuum does not have to be a wall or a membrane. It can also be a process that changes the quality of the subject of study. Chemical reactions as well as processes of proliferation and evolution are examples of this.

In thermodynamics systems are classified as follows according to the nature of their boundary against their environment:

- The *isolated system*: this is an idealized system that does not exchange any kind of energy or matter with its environment;
- The *closed system*: this system can exchange all kinds of energy with its environment, but not matter;
- The *open system*: it can exchange both energy and matter with its environment.

The closed system can be influenced by its environment, and can cause changes in its environment. However, it cannot be involved in an exchange of matter.

The state of a system can be described by a number of *state variables*. These are either extensive or intensive parameters. *Intensive* parameters are nonadditive and independent of the size of the system (e.g., temperature, concentration, pressure, density). Conversely *extensive* parameters are additive when two systems are combined (e.g., mass, volume).

Changes in a system are often characterized by differentials of its state variables. A *differential* describes a very small change of a dependent variable (dy), if in a function $y = f(x)$, a small change in the variable (dx) occurs. It can be calculated from the product of the first derivative of the function $f(x)$, multiplied by dx:

$$\mathrm{d}y = f'(\mathrm{d}x) \tag{3.1}$$

Most thermodynamic equations are functions with several variables. Hence, the derivatives can be obtained with respect to one variable if the others are kept constant.

This procedure is called *partial differentiation*. It has a special notation with the parameters that are to be kept constant put as subscript to the symbols in parentheses. The following example is quoted out of context, to demonstrate this.

$$\left(\frac{\partial G}{\partial n_i}\right)_{p,T,n_j} = \mu_i \quad \text{for} : j \neq i \tag{3.2}$$

The partial derivative of the Gibbs free energy G with respect to the molar number of substance i, when pressure (p), temperature (T), and the molar number of all the other substances (n_j) are kept constant, gives per definition, the chemical potential (μ_i) of the substance i. In general, this is the same procedure as is used when a function with several dependent variables is represented graphically in a two-dimensional plot against one selected variable, keeping all other variables constant.

Small changes in a state function with several variables can be represented by a so-called *total differential*. For this, all partial differentials of this function must be summarized. These partial differentials are calculated as shown in Eq. 3.1, using,

however, partial derivatives. The following equation for example, would apply to the Gibbs free energy [$G(p, T, n_i)$]:

$$\mathrm{d}G = \left(\frac{\partial G}{\partial p}\right)_{T,n_j} \mathrm{d}p + \left(\frac{\partial G}{\partial T}\right)_{p,n_j} \mathrm{d}T + \sum_{i=1}^{m} \left(\frac{\partial G}{\partial n_i}\right)_{p,T,n_j} \mathrm{d}n_i \tag{3.3}$$

The mathematical definition of the total differential is of very great physical importance to thermodynamics. This will be indicated by the following chain of statements with reversible logical connections:

dG is a total differential	$\leftrightarrow$	G is a state function	$\leftrightarrow$ G depends only on the state of the system, and not on the way in which that state was achieved

For this reason it is important that this property of a function is able to be mathematically proven.

A differential equation is not always written in the easily followed way shown in Eq. 3.3.

Often it is presented as a *Pfaffian differential equation:*

$$\mathrm{d}n = L\mathrm{d}x + M\mathrm{d}y + N\mathrm{d}z \tag{3.4}$$

The capital letters here represent any variable. There is a mathematical indicator which allows one to determine whether dn is a total differential. This is the so-called *Cauchy condition*, stressing that dn is a total differential if the following conditions are fulfilled:

$$\frac{\partial L}{\mathrm{d}y} \overset{!}{=} \frac{\partial M}{\partial x}; \qquad \frac{\partial M}{\mathrm{d}z} \overset{!}{=} \frac{\partial N}{\partial y}; \qquad \frac{\partial L}{\mathrm{d}z} \overset{!}{=} \frac{\partial N}{\partial x} \tag{3.5}$$

If this is applied to Eq. 3.3, this means:

$$\frac{\partial\left(\frac{\partial G}{\partial p}\right)}{\partial T} = \frac{\partial\left(\frac{\partial G}{\partial T}\right)}{\partial p} = \frac{\partial^2 G}{\partial T\,\partial p} \tag{3.6}$$

From Eq. 3.6 it follows additionally that for such functions if they are differentiated several times, the sequence of differentiations is unimportant. We will use this property in a later derivation (Eq. 3.82).

Total differentials not only result from energetic parameters. This formalism of course can be applied to any state function. Volume changes in mixed phases for example, can be described by the following total differential equation:

$$\mathrm{d}V = \bar{V}_1\mathrm{d}n_1 + \bar{V}_2\mathrm{d}n_2 + \bar{V}_3\mathrm{d}n_3 + \ldots + \bar{V}_m\mathrm{d}n_m \tag{3.7}$$

where dV is the shift of the volume that occurs when the molar number of one or more components of the system is changed. Furthermore, Eq. 3.7 allows one to define the partial molar volume of a given substance i:

$$\bar{V}_i = \left(\frac{\partial V}{\partial n_i}\right) \quad \text{for} : j \neq i \tag{3.8}$$

The partial molar volume has the inverse unit as the concentration, namely: $m^3\ mol^{-1}$.

3.1.2 Gibbs Fundamental Equation

The scientific basis of thermodynamics is its three principles which are founded on experimentally verifiable, empirical facts. Upon this solid foundation a framework of definitions and relations has been built up which enables far-reaching postulations on all kinds of energy transformations to be made.

The principle of conservation of energy, the so-called *first law of thermodynamics* states that there must exist a physical parameter having the property of a state function, which includes the consequences discussed in Sect. 3.1.1. Work (W), as a physical parameter does not comply with this condition. General experience shows that a change of a system from state A to state B can be achieved in many ways that differ greatly from one another in the amount of work that is required. Therefore, work cannot be a state function that could be used to characterize the energy state of a system independently of the way in which it was achieved.

Let us introduce a parameter called *internal energy* (U). It shall be defined as a state function which, as such, has a total differential dU. Let furthermore the internal energy of a system be increased by dU if a certain amount of heat (dQ) is introduced into the system, and/or if certain work (dW) is done in the system:

$$\mathrm{d}U = \mathrm{d}Q + \mathrm{d}W \tag{3.9}$$

This equation contains the essence of the first principle of thermodynamics. Usually, it is derived in detail with the help of the so-called *Carnot cycle* reflecting processes in steam engines. For simplification, not losing any detail of biophysical relevance, we choose here the simpler way of defining this.

Both, the differentials dQ, as well as dW, are reflections of a change of energy. However, according to the *second principle of thermodynamics*, the heat (Q) differs from all other forms of energy because it possesses a specific property: Any form of energy can be completely transformed into heat but heat itself can only partly be transformed into work.

Here again the entropy (S) has to be inserted. It is the same parameter as introduced in Sect. 2.1.1 in its statistical character (Eq. 2.4). In phenomenological

thermodynamics entropy appears as a kind of measure of heat quality. For a quasi-reversible process it is defined as follows:

$$\mathrm{d}S = \frac{\mathrm{d}Q_{\mathrm{rev}}}{T} \tag{3.10}$$

This equation offers an expression for $\mathrm{d}Q$ which can be introduced into Eq. 3.9.

Let us now consider in more detail the differential of work ($\mathrm{d}W$) in Eq. 3.9. This can be a sum of different kinds of work. Each being a product of a work coefficient multiplied by a work coordinate. *Work coefficients* are intensive parameters indicating a kind of measure of the performed work. In contrast, *work coordinates* are extensive parameters, reflecting the effort of the performed work. For example, the work ($\mathrm{d}W_p$) which is done when a gas is compressed by a pressure p resulting in a volume alteration $\mathrm{d}V$ will be:

$$\mathrm{d}W_p = -p\,\mathrm{d}V \tag{3.11}$$

where p represents the work coefficient and $\mathrm{d}V$ the work coordinate. The sign of this equation depends on the definition of the work differential. A positive $\mathrm{d}W$ means that there is an increase in the work done in favor of the system. In this case work is achieved through the compression, i.e., a negative differential of the volume.

Work can be done in many different ways. An expansion of a material, for example, means elongation (dl) in response to the application of a force ($\mathbf{F}$):

$$\mathrm{d}W_l = \mathbf{F}\,\mathrm{d}l \tag{3.12}$$

A cell can do work by transporting a certain number of atoms or molecules ($\mathrm{d}n$) against a concentration gradient. At this point, the chemical potential (μ) must be introduced as the work coefficient. We will come back to this parameter in detail later (Eq. 3.33). In this case the work differential can be written as follows.

$$\mathrm{d}W_n = \mu\ \mathrm{d}n \tag{3.13}$$

Let us finally, from among the many other possible examples, consider charge transport. If a particular amount of charge ($\mathrm{d}q$) is transported against an electric potential (ψ), then the electrical work done will be:

$$\mathrm{d}W_q = \psi\ \mathrm{d}q \tag{3.14}$$

Equations 3.11, 3.12, 3.13, and 3.14 can be combined:

$$\mathrm{d}W = -p\ \mathrm{d}V + \mathbf{F}\ \mathrm{d}l + \mu\ \ \mathrm{d}n + \psi\ \mathrm{d}q \tag{3.15}$$

Considering that usually in the system a number of m substances are transported, then this equation can be expanded as follows:

$$\mathrm{d}W = -p\ \mathrm{d}V + \mathbf{F}\ \mathrm{d}l + \sum_{i=1}^{m} \mu_i\ \mathrm{d}n_i + \psi\ \mathrm{d}q \tag{3.16}$$

This is a more detailed form of the work differential which together with Eq. 3.10, can be introduced into Eq. 3.9:

$$\mathrm{d}U = T\mathrm{d}S - p\ \mathrm{d}V + \mathbf{F}\mathrm{d}l + \sum_{i=1}^{m} \mu_i\ \mathrm{d}n_i + \psi\ \mathrm{d}q \tag{3.17}$$

Equation 3.17 is a differential form of the so-called *Gibbs fundamental equation*. Of course it can be expanded optionally by adding more kinds of work differentials, for example for magnetic field influences (see: Eq. 4.19, Sect. 4.4). Alternatively, this equation will be automatically reduced if certain changes become irrelevant. For example, suppose a defined transformation within a system is not accompanied by a mechanical strain. Then l remains constant, and consequently, $\mathrm{d}l = 0$. As a consequence, the corresponding term disappears from the equation.

In Eq. 3.17 the Gibbs fundamental equation appears in the form of a *Pfaffian differential*. Such expressions can be integrated under certain conditions, which apply in this case. This gives:

$$\mathrm{U} = T\ S - p\ V + \mathbf{F}l + \sum_{i=1}^{m} \mu_i\ n_i + \psi\, q \tag{3.18}$$

It must be noted that the transition from Eq. 3.17 to 3.18 does not mean a simple elimination of the differential operators; it is the result of a proper integration which is not described here!

Using the rule given in Eq. 3.3, this process of integration can be reversed. It gives:

$$\begin{aligned}\mathrm{d}U = {} & T\ \mathrm{d}S + S\ \mathrm{d}T - p\ \mathrm{d}V - V\ \mathrm{d}p + \mathbf{F}\,\mathrm{d}l + l\ \mathrm{d}\mathbf{F} \\ & + \sum_{i=1}^{m} \mu_i\ \mathrm{d}n_i + \sum_{i=1}^{m} n_i\ \mathrm{d}\mu_i + \psi\ \mathrm{d}q + q\ \mathrm{d}\psi\end{aligned} \tag{3.19}$$

A comparison of this result with the initial equation (3.17) shows that the following condition must be satisfied:

$$S\ \mathrm{d}T - V\ \mathrm{d}p + l\ \mathrm{d}\mathbf{F} + \sum_{i=1}^{m} n_i\ \mathrm{d}\mu_i + q\ \mathrm{d}\psi = 0 \tag{3.20}$$

This is the so-called *Gibbs–Duham equation*. It is useful for some calculations because it allows one to reduce the degree of freedom of a system by one variable.

It has proved useful to define not only the internal energy (U), but also three further energy functions. In some books the introduction of these parameters is explained in a physical way, discussing processes of vapor compression, etc., but it seems to be simpler just to accept the definitions of these parameters, and subsequently substantiate their usefulness.

The definitions are:

$$\text{enthalpy:} \quad H = U + pV \tag{3.21}$$

$$\text{Helmholtz free energy} \quad F = U - TS \tag{3.22}$$

$$\text{Gibbs free energy} \quad G = H - TS \tag{3.23}$$

The Gibbs fundamental equation (Eq. 3.17) can now easily be written down for these new defined functions. Let us first transform Eq. 3.21 into a total differential, according to Eq. 3.3. Using the definition (Eq. 3.21), the enthalpy (H) is a function of: U, p, and V. This gives:

$$\mathrm{d}H = \left(\frac{\partial H}{\partial U}\right)_{p,V} \mathrm{d}U + \left(\frac{\partial H}{\partial p}\right)_{U,V} \mathrm{d}p + \left(\frac{\partial H}{\partial V}\right)_{U,p} \mathrm{d}V \tag{3.24}$$

From Eq. 3.21 follows directly:

$$\left(\frac{\partial H}{\partial U}\right)_{p,V} = 1; \qquad \left(\frac{\partial H}{\partial p}\right)_{U,V} = V; \qquad \left(\frac{\partial H}{\partial V}\right)_{U,p} = p \tag{3.25}$$

which, when combined with Eq. 3.24, results in

$$\mathrm{d}H = \mathrm{d}U + V\mathrm{d}p + p\mathrm{d}V \tag{3.26}$$

Combining this with Eq. 3.17 gives the Gibbs fundamental equation for dH:

$$\mathrm{d}H = T\mathrm{d}S + V\ \mathrm{d}p + \mathbf{F}\ \mathrm{d}l + \sum_{i=1}^{m} \mu_i\ \mathrm{d}n_i + \psi\ \mathrm{d}q \tag{3.27}$$

In the same way it is possible to derive from Eq. 3.22 the relations:

$$\mathrm{d}F = S\ \mathrm{d}T - p\ \mathrm{d}V + \mathbf{F}\ \mathrm{d}l + \sum_{i=1}^{m} \mu_i\ \mathrm{d}n_i + \psi\ \mathrm{d}q \tag{3.28}$$

and:

$$\mathrm{d}G = S\ \mathrm{d}T + V\ \mathrm{d}p + \mathbf{F}\ \mathrm{d}l + \sum_{i=1}^{m} \mu_i\ \mathrm{d}n_i + \psi\ \mathrm{d}q \tag{3.29}$$

The choice whether Eq. 3.17, 3.27, 3.28, or 3.29 should be used to calculate a particular system depends on the external conditions and the question which is being asked.

Investigating a system under isobaric conditions ($p = \text{const.}$, i.e., $\mathrm{d}p = 0$), it is useful to apply the equation for $\mathrm{d}H$ (Eq. 3.27), or for $\mathrm{d}G$ (Eq. 3.29), because in this case the term of the volume expansion work ($V\mathrm{d}p$) vanishes. This corresponds to the situation of most biological investigations. Therefore we will use mostly the enthalpy (H) and the Gibbs free energy (G) in all further biophysical calculations instead of the inner energy (U) and the Helmholtz free energy (F).

If the conditions are isothermal ($\mathrm{d}T = 0$), as well as isobaric ($\mathrm{d}p = 0$), then in Eq. 3.29 the term, connected with heat, as well as that for volume work will vanish. Hence, $\mathrm{d}G$ expresses directly the deviation of the energy content, as a result of work which was done. The gradient of free Gibbs energy therefore indicates the direction of a spontaneous process in the same way as a gradient of the potential energy indicates the path of a rolling sphere on an inclined surface.

All these forms of the Gibbs fundamental function (Eqs. 3.27, 3.28, 3.29), as well as Eq. 3.7 for partial volume, can be integrated according to Eq. 3.17 for $\mathrm{d}U$.

In this way, a particular chemical reaction can be characterized by differences of these parameters:

$$\begin{aligned} \Delta_R G &= G_{\text{product}} - G_{\text{substrat}} \\ \Delta_R H &= H_{\text{product}} - H_{\text{substrat}} \\ \Delta_R S &= S_{\text{product}} - S_{\text{substrat}} \end{aligned} \tag{3.30}$$

The parameter ΔG_R indicates whether this reaction occurs spontaneously, i.e., whether $G_{\text{product}} < G_{\text{substrate}}$, i.e., $\Delta G_R < 0$. Conversely the reaction enthalpy ΔH_R is a measure of the thermal characteristics of isobaric processes. This means:

$\Delta_R H > 0$ – endothermic reaction
$\Delta_R H < 0$ – exothermic reaction

How does one obtain the parameters as used in Eq. 3.30? Only in the case of entropy are absolute values available. This is a result of the *third principle of thermodynamics*, based on the heat theorem of Walther Nernst, stating that at a temperature of absolute zero the entropy becomes zero.

In contrast to entropy, no absolute values exist for the energetic parameters. They always need a defined reference value. Therefore, *standard energies of formation* ($\Delta_F U$, $\Delta_F H$, $\Delta_F F$, and $\Delta_F G$) are defined as energetic changes that occur when a substance is formed from its elements under standard conditions ($T = 297$ K), or better: *would occur*, because in most cases a direct synthesis of the substance from its elements is impossible. In this case, they are estimated from particular sets of chemical reactions starting from substances with known energy of formation. This is possible using the definition of these parameters as state functions, their amount being independent of the way the state was achieved (see definition in Sect. 3.1.1).

According to definition (2.23), for isothermal systems this is:

$$\Delta_R G = \Delta_R H - T\ \Delta_R S \tag{3.31}$$

A spontaneous reaction ($\Delta_R G < 0$) therefore requires an exothermic reaction ($\Delta_R H < 0$) or the condition $\Delta_R H < T\Delta_R S$. In this case the direction of the reaction is determined by the rise in entropy. This type of process is called an *entropy-driven reaction*. The classical example of such a reaction is the melting of ice. In Sect. 2.2.2 we already discussed various biomolecular reactions of this type in context with the structure of water.

Let us now consider the chemical potential which we need for numerous considerations in the following sections. The chemical potential (μ_i) of the substance i is particularly important for the following calculations. It can be easily defined using Eqs. 3.17, 3.27, 3.28, or 3.29:

$$\mu_i = \left(\frac{\partial U}{\partial n_i}\right)_{S,V,l,n_j,q} = \left(\frac{\partial H}{\partial n_i}\right)_{S,p,l,n_j,q} = \left(\frac{\partial F}{\partial n_i}\right)_{T,V,l,n_j,q} = \left(\frac{\partial G}{\partial n_i}\right)_{T,p,l,n_j,q} \tag{3.32}$$

The chemical potential therefore, is a partial expression having the dimensions J mol^{-1}. The chemical potential of the substance i (μ_i) can be calculated from the concentration (c_i), resp. the chemical activity (a_i) of this substance using:

$$\mu_i = \mu_i^0 + RT\ \ln a_i \tag{3.33}$$

The chemical activity (a_i) is a kind of effective concentration. Its relation to concentration is given by:

$$a_i = f_i c_i \tag{3.34}$$

In this equation, f_i is the *coefficient of activity*. In ideal solutions, $f_i = 1$, that is, the activity of the dissolved substance is equal to its concentration c_i. Usually f_i decreases as the concentration in the solution increases (see Fig. 3.1). For dissociating salts, f represents an average activity coefficient for the ions. For example, the ions in a 100-mM solution of NaCl show an average activity coefficient of 0.8. The chemical activity of this solution therefore is equal to an ideal solution with a concentration of only 80 mM. In contrast to the coefficient of activity, which is a dimensionless number, the chemical activity has the same units as the concentration.

In some cases it may be useful to employ the mole fraction as a measure of concentration. The *mole fraction* of a substance i is defined as the number of moles of that substance (n_i), divided by the total number of moles of all substances present:

$$x_i = \frac{n_i}{\sum_{i=1}^{m} n_i} \tag{3.35}$$

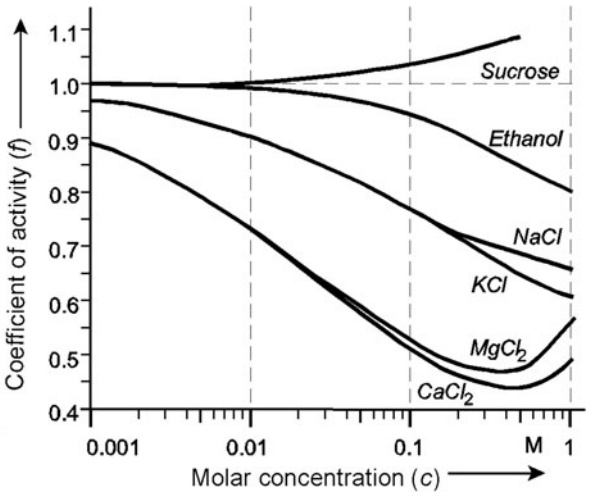

Fig. 3.1 Coefficients of activity (f_i) of various substances as functions of their concentrations (c_i) in aqueous solutions under standard conditions ($T = 297$ K)

where:

$$\sum_{i=1}^{m} x_i = 1 \tag{3.36}$$

According to Eq. 3.34, the mole fraction of a substance can also be expressed as the mole fraction activity (a_{xi}).

The standard potential (μ_i^0) can be easily defined by means of Eq. 3.33. It follows from this equation that when $a_i = 1$, $\mu_i = \mu_i^0$. The standard potential therefore is the chemical potential of a substance i in a solution, with an activity of 1 M, if this measure of concentration is applied. Using in Eq. 3.33 as concentration measure the mol fraction (x_i) of a substance, or its mol fraction activity (a_{xi}), then the chemical standard potential (μ_i^0) is determined by the chemical potential of the pure substance ($a_{ix} = 1$).

A further extension of the Gibbs fundamental equation concerns the term $\psi \mathrm{d}q$. The charge of a single charged ion is determined by the Faraday constant (F). The charge on n moles of a z-fold charged ion is obtained as follows:

$$q = zn\mathrm{F} \tag{3.37}$$

This is a function with one independent variable (n). Therefore, it is easily transformed into a differential according to Eq. 3.1:

$$\mathrm{d}q = \left(\frac{\mathrm{d}q}{\mathrm{d}n}\right)\mathrm{d}n = z\mathrm{F}\ \mathrm{d}n \tag{3.38}$$

If more than one ion is in the solution, the charges can be summarized:

$$\mathrm{d}q = \sum_{i=1}^{m} z_i \mathrm{F}\ \mathrm{d}n_i \tag{3.39}$$

Introducing this expression into the Gibbs fundamental equation, the equation then gets two terms with the differential dn_i. It is obviously useful to combine these terms. First, consider these two terms of the Gibbs equation in isolation from the other terms of this equation:

$$\psi \sum_{i=1}^{m} z_i \mathrm{F}\ \mathrm{d}n_i + \sum_{i=1}^{m} \mu_i\ \mathrm{d}n_i = \sum_{i=1}^{m} (\mu_i + z_i\,\mathrm{F}\ \psi)\mathrm{d}n_i = \sum_{i=1}^{m} \tilde{\mu}_i\ \mathrm{d}n_i \tag{3.40}$$

Here, the expression inside the brackets taken together is described as the *electrochemical potential* ($\tilde{\mu}_i$) of the substance i.

$$\tilde{\mu}_i = \mu_i + z_i\,\mathrm{F}\,\psi \tag{3.41}$$

The electrochemical potential is the basis for most electrochemical calculations, and thus forms an important starting point for further considerations.

3.1.3 Force and Motion

After introduction of the generalized functions for the energetic state of a system in the previous section we will now consider their application in determining forces leading to any sort of motion.

A sphere is rolling downhill. It is moving spontaneously from a position with a higher potential energy to one with a lower potential. The direction of this movement follows a force vector ($\mathbf{X}$) and is, consequently, determined by the negative gradient of the energy U.

$$\mathbf{X} = -\mathrm{grad}\ \ U \tag{3.42}$$

If consideration of the energy gradient is confined to the direction of the x-coordinate, this equation can be simplified to give:

$$\mathbf{X}_x = -\frac{\mathrm{d}U}{\mathrm{d}x}\mathbf{i} \tag{3.43}$$

where $\mathbf{i}$ is simply a unit vector, i.e., a vector with the amount of 1, and an arrow, directing toward the x-coordinate. For $\mathrm{d}U$, any appropriate energy state function should be substituted as shown in Sect. 3.1.2. Let us consider for example the force acting on a charge (q) in an electric field in the x-direction ($\mathbf{E} = -\mathrm{d}\psi/\mathrm{d}x{\cdot}\mathbf{i}$). Substituting for $\mathrm{d}U$ in Eq. 3.43, the expression, according to Eq. 3.17, gives:

$$\mathbf{X}_q = -\frac{\mathrm{d}U}{\mathrm{d}x}\mathbf{i} = -\frac{\mathrm{d}\psi}{\mathrm{d}x}\mathbf{i}\ q = \mathbf{E}\ q \tag{3.44}$$

This assumes that there is no other gradient in the x-direction, i.e., that neither p, nor T, nor μ are functions of x. The introduction of the field strength ($\mathbf{E}$) is in accordance with the definition given in Eq. 2.45. Equation 3.44 is identical with Eq. 2.46, which was derived by other approaches.

Equation 3.44 cannot be applied to practical calculations of ion transport because only the transport of an electrical charge is considered. In contrast to the movement of an electron, the transport of ions always means that there is an additional change in concentration. For transport of a noncharged substance (i), the negative gradient of its chemical potential (μ_i) is the driving force:

$$\mathbf{X}_i = -\text{grad } \mu_i \tag{3.45}$$

Ions, in contrast, are driven by the gradient of the electrochemical potential ($\tilde{\mu}$), which according to Eq. 3.41 includes the electric potential. Applying the differential operator, the electric potential (ψ) transforms into an electric field strength ($\mathbf{E}$):

$$\mathbf{X}_i = -\text{grad } \tilde{\mu}_i = -(\text{grad } \mu_i - z_i \, \text{F}\mathbf{E}) \tag{3.46}$$

There are many kinds of movement in biological systems that are calculated by biophysical approaches. Their range covers electron transfer, structural changes of molecules, chemical reactions, fluxes of molecules and ions, streaming of liquids in the body, and finally, mechanical movements of limbs and whole bodies. Fluxes occupy a central position in the study of movements in biology, and therefore, these will be considered here for the sake of simplicity, as a sort of generalized movement.

The flux ($\mathbf{J}_i$) is defined as the amount of a substance (i) that passes perpendicularly through a unit of surface per unit of time. This definition shows that the flux in general is a vector. Often the flux through a membrane is considered, the direction of which is always predicted.

The relation between flux $\mathbf{J}_i$ of component i, and its velocity $\mathbf{v}_i$ is:

$$\mathbf{J}_i = c_i \mathbf{v}_i \tag{3.47}$$

A system which is traversed by a flux, and where no substance is added or removed to this flow, is called a *conservative* one. In this case the following conditions apply:

$$\text{div } \mathbf{J}_i = 0 \tag{3.48}$$

The differential operator div ("divergence") can be replaced by the Nabla operator (∇), as explained in Sect. 2.2.1. In contrast to the operator "grad," which, applied to a scalar results in a vector, the "div" operator is to be applied to a vector, producing a scalar. The Nabla operator (∇) is applicable to differentiate vectors as well as scalars.

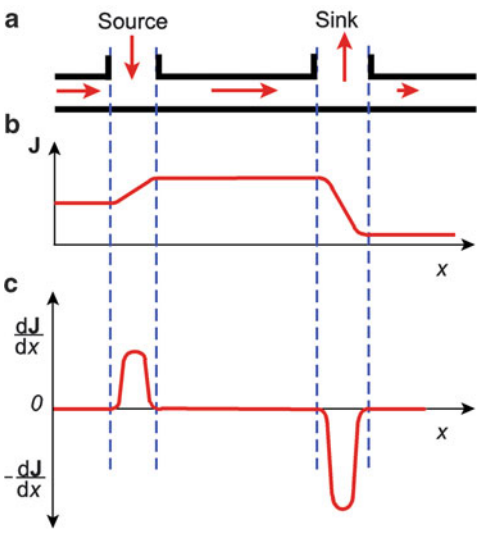

Fig. 3.2 (**a**) graphic representation of a linear flow with a source, and a sink; (**b**) the flux (**J**) as a function of x; (**c**) the change of the flow ($d\mathbf{J}/dx$) as a function of x

If $\mathrm{div}\mathbf{J}_i > 0$ in such a system which is being traversed by a flux, it indicates that there is a source that is adding substance to the flux. However, if $\mathrm{div}\mathbf{J}_i < 0$, then there will be a removal of some of the substance from the flux like a sink. Figure 3.2 illustrates this situation using an example where the flow is simply in the x-direction.

This formalism is applied, describing for example fluxes through biological tissue. For the transport of ions, this system is conservative, because no accumulation or depletion of ions occurs. In contrast, for the transport of oxygen the condition $\mathrm{div}\mathbf{J}_o < 0$ holds, because the tissue uses oxygen for its own respiration.

If a constant force acts on a body, then the latter accelerates. However, as the velocity of this body increases, friction is likely to increase too. When both the driving force and the frictional force become the same amount, then the body will move with a constant velocity. This is a special case of a stationary state, the so-called *stationary motion*.

It is a fundamental experience in physics that the particular relation between force and movement characterizes regions with different qualities. If, for example, a comparatively minor force acts on a body, then the body will attain a velocity that is simply proportional to the force; this is a linear force-movement relationship and therefore the linear approaches of irreversible thermodynamics are applicable. If the same body is more forcibly moved, then the frictional force will increase in a stronger way, the frictional coefficient is no longer constant, and a nonlinear force-movement approach is necessary. Therefore, nonlinear thermodynamics, far from equilibrium, must be applied including some new qualitative properties.

This concept, illustrated here by a mechanical example, has a more general application. The linear approach can be formulated in a generalized form as an equation of motion in the following way:

$$\mathbf{J}_i = L_i\, \mathbf{X}_i \tag{3.49}$$

where the coefficient L_i is a kind of *generalized conductance*. In the same way, the following equation can be written:

$$\mathbf{X}_i = R_i \mathbf{J}_i \tag{3.50}$$

In this case a *resistance factor* is applied: $R_i = 1/L_i$. Ohm's law is a special form of this general expression:

$$U = R\,I \tag{3.51}$$

where U in this case is the electromotive force, depending on the potential gradient (gradψ), and I is the electric current.

In a similar way force and velocity can be related in a linear way. For this it is necessary to introduce a mobility factor (ω), and its corresponding *coefficient of friction* (f):

$$\mathbf{v} = \omega \mathbf{X} = \frac{\mathbf{X}}{f} \tag{3.52}$$

The flux equation (3.47) can now be written as follows:

$$\mathbf{J}_i = c_i\,\omega_i \mathbf{X}_i \tag{3.53}$$

We introduced the Stoke's equation (2.34) for the discussion of molecular movement in Sect. 2.1.6. This is also a particular expression of this linear approach. It indicates the frictional force ($\mathbf{F}$) of a sphere with radius (r), moving in a liquid with viscosity (η) at velocity ($\mathbf{v}$). Under conditions of stationary motion, this friction force is equal to the driving force:

$$\mathbf{F} = 6\pi r \eta \mathbf{v} \tag{3.54}$$

In view of Eq. 3.52, the following analogy is obvious:

$$f = \frac{1}{\omega} = 6\pi\, r\, \eta \tag{3.55}$$

The definitive mechanical behavior of bodies moving in water will be discussed in detail in Sect. 3.7. Here it is only being used as an example for the consideration of general principles.

Figure 3.3 shows the frictional coefficient (f) for a sphere with a diameter of 1 cm in water as a function of its velocity ($\mathbf{v}$). The frictional coefficient only remains constant up to a velocity of about 1 mm s^{-1}, corresponding to Eq. 3.55. As the velocity increases above this, the frictional factor deviates, at first slightly, and then greatly from the horizontal line. This illustrates the transition of the system from linear to nonlinear behavior. This is the transition to far-equilibrium conditions, where nonlinear thermodynamic approaches must be applied.

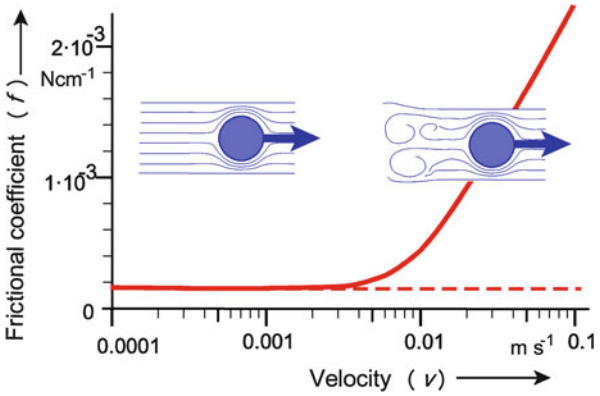

Fig. 3.3 The coefficient of friction ($f = \mathbf{F}/\mathbf{v}$) of a sphere with a diameter of 0.01 m in water as a function of its velocity ($\mathbf{v}$)

Let us explain the genesis of such a nonlinear relation by a simple example. Let a phenomenological coefficient be a sum of a constant term (L_i), and a variable term L_i', which is proportional to the flux ($\mathbf{J}_i$). In this case, the linear approach of Eq. 3.49 transforms into:

$$\mathbf{J}_i = (L_i + L_i' \mathbf{J}_i)\, \mathbf{X}_i \tag{3.56}$$

Solving this equation for $\mathbf{J}_i$, gives:

$$\mathbf{J}_i = \frac{L_i \mathbf{X}_i}{1 - L_i' \mathbf{X}_i} \tag{3.57}$$

Hence we have a nonlinear function $\mathbf{J}_i(\mathbf{X}_i)$.

The qualitative consequences of these nonlinear approaches will be considered further in the next section. Let us at present remain in the field of linear thermodynamics.

One of the fundamental statements of linear thermodynamics concerns the coupling of forces and movements in a system. If different kinds of motions and forces occur simultaneously in a system, they will influence each other.

Let us consider again the flux ($\mathbf{J}_i$) as a generalized kind of motion. Nonequilibrium thermodynamics in its scope of linear approaches allows us to formulate a set of phenomenological equations forming a flux matrix. This is the mathematical expression of the general statement whereas all fluxes in a system in principle are coupled with each other.

The simple Eq. 3.49, therefore, will be expanded to the following set of equations:

$$\begin{aligned}
\mathbf{J}_1 &= L_{11}\mathbf{X}_1 + L_{12}\mathbf{X}_2 + L_{13}\mathbf{X}_3 + \ldots + L_{1n}\mathbf{X}_n \\
\mathbf{J}_2 &= L_{21}\mathbf{X}_1 + L_{22}\mathbf{X}_2 + L_{23}\mathbf{X}_3 + \ldots + L_{2n}\mathbf{X}_n \\
\mathbf{J}_3 &= L_{31}\mathbf{X}_1 + L_{32}\mathbf{X}_2 + L_{33}\mathbf{X}_3 + \ldots + L_{3n}\mathbf{X}_n \\
&\ldots \\
\mathbf{J}_n &= L_{n1}\mathbf{X}_1 + L_{n2}\mathbf{X}_2 + L_{n3}\mathbf{X}_3 + \ldots + L_{nn}\mathbf{X}_n
\end{aligned} \tag{3.58}$$

In these equations the vector notation of fluxes and forces is still retained (bold letters), regardless of whether they may in fact, in rare cases, be scalars, which we will discuss in the following.

The parameters L_{mn} are phenomenological coefficients, also called *coupling coefficients*, *cross coefficients,* or *Onsager coefficients* (this approach was introduced first by Lars Onsager in 1931). In reality, this general set of equations may be reduced, because a flux $\mathbf{J}_m$ is coupled with a force $\mathbf{X}_m$ only when $L_{mn} \neq 0$.

Equation 3.58 shows that n forces with their corresponding fluxes require a set of equations with n^2 coupling coefficients. Onsager, however, was able to show that this matrix is symmetric. This means that near equilibrium the following relation holds:

$$L_{mn} = L_{nm} \qquad \text{for}: \quad n \neq m \tag{3.59}$$

This is *Onsager's law on the reciprocal relation*. It leads to a significant reduction of the coefficients in the matrix from n^2, down to $n(n + 1)/2$.

Directly linked pairs of forces and fluxes, as for example $\mathbf{J}_1$ and $\mathbf{X}_1$, $\mathbf{J}_2$ and $\boldsymbol{X}_2$, . . ., $\mathbf{J}_n$ and $\mathbf{X}_n$, are called *conjugated*. The coefficients (L_{nn}), linking these pairs, are always positive. If two fluxes really are coupled, the following condition must hold:

$$L_{mm} \cdot L_{nn} \geq L_{mn}^2 \tag{3.60}$$

From this, a *degree of coupling* (q_{mn}) can be defined:

$$q_{mn} = \frac{L_{mn}}{\sqrt{L_{mm} L_{nn}}} \tag{3.61}$$

This degree of coupling can vary as follows: $1 \geq q_{mn} \geq 0$. When $\boldsymbol{q_{mn}} = 0$, the fluxes are completely independent of each other, when $q_{mn} = 1$, there is maximal coupling.

As mentioned earlier in Eq. 3.58, all fluxes and forces are printed in bold letters as vector parameters. At the same time we mentioned that we will consider here fluxes in a very general sense, symbolizing all kinds of motion. This means that not only true fluxes of matter which really are vectors going in a particular direction, but for example, also chemical reactions will be considered. The flux, as mentioned in this equation, therefore, can also mean a production of a substance, or the removal of it, by a chemical reaction. In this case however the flux does not remain a vector, but becomes a scalar. How can we include these scalar fluxes into a matrix together with vectors and not violate mathematical rules?

Let us consider a simple set of flux equations, including a transport of a substance ($\mathbf{J}_i$) and a chemical reaction, the rate of which we will denote by the scalar flux J_R. Formal mathematics appears to allow only the following possibility:

$$\begin{aligned} J_R &= L_{RR} X_R + \mathbf{L}_{Ri} \mathbf{X}_i \\ \mathbf{J}_i &= \mathbf{L}_{iR} X_R + L_{ii} \mathbf{X}_i \end{aligned} \tag{3.62}$$

In the first equation, the product of two vectors ($\mathbf{L}_{Ri}\mathbf{X}_i$) gives a scalar, as well as the product of scalars ($L_{RR}X_R$), and so, this equation is a sum of scalars resulting in a scalar flux (J_R). In contrast, in the second equation all the terms of the sum are vectors. In both equations therefore, the mathematical requirement for homogeneity has been satisfied.

But: what is the meaning of a vectorial coupling coefficient? Introducing this parameter, we declared it as a sort of conductivity. What does a conductivity vector mean? In fact, in so-called *anisotropic systems* vectorial coefficients can appear. In an isotropic system, for example, in an aqueous solution, the mobility of an ion in all directions is equal. The parameter L, therefore is a scalar. Considering, however, the same ion in a pore of a membrane, its movement is possible only in a predetermined direction, and its conductivity consequently becomes a vector.

These circumstances are considered in the so-called *Curie–Prigogine principle*. It states that the direct coupling of scalar and vectorial fluxes is possible only in anisotropic systems. This principle, for example, is important in biophysical considerations of active ion transport. In this case a hydrolysis of ATP is coupled to the transport of an ion against an electrochemical gradient.

This example already indicates that the concept of coupled movements is not limited to mechanical frictional interactions. This in fact is the case in some electro-osmotic processes as described in Sect. 2.3.5 (Fig. 2.45). Furthermore, in Sect. 3.3.1 we will use this approach to introduce Staverman's reflection coefficient, governing the coupling of the osmotic fluxes of water and solute (Eq. 3.150). In general, however, the concept of Onsager coefficients is applied in various networks of biochemical processes.

Further Reading

Katchalsky and Curran 1965; Prigogine 1967; Schnakenberg 1981; Kjelstrup and Bedeaux 2008.

3.1.4 Entropy, Stability, and Stationary States

The second principle of thermodynamics states that an isolated system moves spontaneously towards a maximum in its entropy. When this state is achieved, then the system is in thermodynamic equilibrium. In the same way, the decrease of the free energy down to a minimum can be considered as the way towards the equilibrium in the sense of the second principle.

Any movement as a result of energy transformation leads to an increase in the entropy of the system or its environment. The term *entropy production* ($\sigma = \mathrm{d}S/\mathrm{d}t$) has been introduced to characterize this process. The entropy production is always positive, but can approach zero. The condition: $\sigma = 0$ would mean an idealized reversible process. Thermodynamically, a process is defined as being reversible if it can be repeated an arbitrary number of times without requiring the supply of additional energy.

To prevent misunderstanding, the different meanings of the term "reversible" in physics, chemistry, and biology must be pointed out. In physics, the term "reversible" is used according to the thermodynamic definition, i.e., connected with the above-mentioned condition: $\sigma = 0$. When a chemist speaks about a "reversible reaction," or a "reversible electrode," he only means processes that in principle could run in both directions, independently of the required energy. Finally, the biologist states that a change in a biological system is "reversible" when it is able to reverse an induced change so that no irreparable damage is caused (for example: reversible inhibition of metabolic processes).

Let us consider the total entropy balance of a system. In closed systems, as a result of energy transformations, only an entropy increase is possible up to the point where the thermodynamic equilibrium is established without any further entropy-producing processes. In open systems, however, which are able to exchange not only energy but additionally matter, the entropy may change in both directions. An entropy decrease can occur if substances with low entropy content are incorporated, in exchange for entropy-rich substances that are being extruded. To characterize this process, an entropy flux ($\mathbf{J}_S$) is formulated which penetrates the whole system. Hence, the total entropy balance of the system can be written as follows:

$$\frac{\partial S}{\partial t} = -\nabla \mathrm{J}_S + \sigma \tag{3.63}$$

The overall change of the entropy of an open system ($\partial S/\partial t$), therefore, results as the sum of the always positive entropy production (σ), and the divergence of the entropy flux ($\nabla \mathbf{J}_s \equiv \mathrm{div}\ \mathbf{J}_s$), penetrating the system. In reference to the definition in Sect. 3.1.3 (see Fig. 3.2), the system in fact is not conservative in relation to this entropy flux. Depending on the relation of the two terms in the sum (Eq. 3.63), the total entropy change ($\partial S/\partial t$) can become positive as well as negative. The control of the term $\nabla \mathbf{J}_S$ can be considered as the work of a Maxwell demon, as described in Sect. 2.1.2 (Fig. 2.2).

For further considerations it may be useful to introduce a thermodynamically based classification of the various kinds of stationary states. We define a *stationary state* as a state where the structure and parameters are time independent. The reasons, leading to this quality can be quite different. The water level of a lake, for example, can be time independent, i.e., constant, either because there is no inflow into the lake, and no outflow, or because inflow and outflow are equal. These two kinds of stationary states can be distinguished by their entropy production. In the first case no energy is required to maintain this state, therefore there is no entropy production, the system is in *thermodynamic equilibrium* ($\sigma \stackrel{!}{=} 0$). In contrast, the lake with exactly the same in- and outflow is in *steady state*. This is a stationary state with entropy production ($\sigma \stackrel{!}{>} 0$). The thermodynamic definition of the steady state is the only possible one. It seems important to emphasize that a steady state cannot be defined by its kinetic properties.

Let us illustrate this statement with an example: Using radionuclides it is possible to demonstrate that human erythrocytes exchange chloride as well as

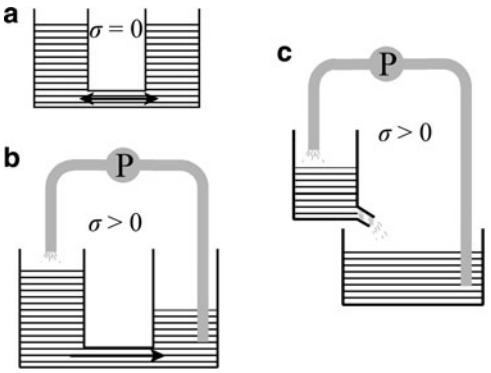

Fig. 3.4 Stationary states in hydraulic models: (**a**) thermodynamic equilibrium ($\sigma = 0$); (**b**) a steady-state system ($\sigma > 0$), which becomes an equilibrium (A), if the pump (P) is stopped; (**c**) a steady-state system ($\sigma > 0$), where stopping the pump (P) would lead to a complete outflow of the liquid from the upper vessel

potassium ions with their environment. With this method, it is possible to measure directly the corresponding exchange rates. This kinetic method of analysis may give the impression that both ions, $^{-}$Cl, as well as $^{+}$K, are in a steady state because in both cases the unidirectional fluxes, outside → in and inside → out, are equal. This, however, is an incorrect conclusion. The chloride ions in fact are distributed passively between the external medium and the cytoplasm, according to their thermodynamic equilibrium. The observed exchange of radioactive chloride results from their self-diffusion by thermal motion. This process is not accomplished by entropy production because no energy is converted. It is like a stochastic exchange of water molecules between the two vessels in Fig. 3.4a. In contrast to this, potassium is pumped actively into the cell by use of metabolic energy against its electrochemical gradient, and diffuses passively back, driven by this gradient (Fig. 3.4b). Both, the active transport as well as the passive diffusion, are true fluxes in the thermodynamic sense, producing entropy. Potassium, therefore, in contrast to chloride, really, is in a stationary state. This example indicates that the above-described kinetic experiment of compartment analyses is unsuitable to distinguish between an equilibrium state and a steady state. We will come back to this problem later in detail (Sects. 3.4.1 and 5.1.1).

As we will see later (Sect. 3.3.3), the steady state of potassium can be transformed into an equilibrium state if the pump is inhibited. In this case, a Donnan equilibrium will be established which means an equilibration of all electrochemical potentials. The steady state of sodium and potassium in the cell therefore resembles case b in Fig. 3.4, passing into the equilibrium state (Fig. 3.4a), when the active transport is inhibited. In contrast to this, various substances do not show the possibility of an equilibrium distribution. If the influx is stopped they disappear completely (Fig. 3.4c).

An important property of all stationary states is their kind of stability. Let us illustrate this by a mechanical example of a sphere on a surface (Fig. 3.5). The requirement for a stationary state, in this case this simply means an equilibrium, is the sphere coming to rest on any small, but horizontal part of the surface. In the case of an *indifferent* state, the surface is horizontal in general. In this case the energy of the sphere will not be changed by alteration of its position. In the case of a *stable*

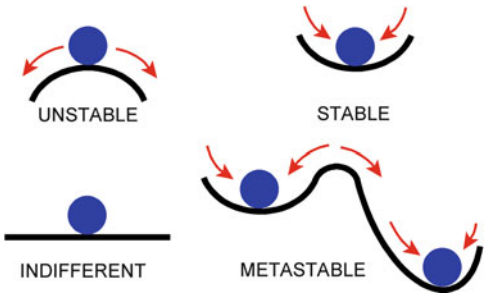

Fig. 3.5 Various kinds of stability conditions

state, every change of the position leads to an increase of the energy of the sphere, and generates a force driving the sphere back to its original state. In contrast an *unstable state* is characterized by a situation where even small changes of the position release forces that cause the system to be deflected even more. Additionally sometimes so-called metastable states are considered. As *metastable*, a stable state can be considered which is delimitated from another one by a small barrier which can easily be overcome.

In the mechanical examples of Fig. 3.5 the shape of the surface automatically indicates the function of its potential energy. In general, however, these surfaces have to be replaced by true functions of the free energy like those of Figs. 2.5 or 2.26.

Figure 3.6 indicates all possible kinds of stationary states. First of all the presence, or the absence of entropy production indicates whether the given stationary state is a *thermodynamic equilibrium* ($\sigma = 0$), or whether it is a *steady state* ($\sigma > 0$). In the case of thermodynamic equilibrium one must distinguish between global and local equilibria. In the case of a *global equilibrium*, the function of free energy indicates only one minimum. This means that no alteration, however strong it may be, can bring the system into another equilibrium state. An example is the equilibrium distribution of an ion between the cell and its environment. In contrast, in the case of the *local equilibrium*, the energetic function indicates two or more minima which are separated by more or less high energy barriers. As an example, isotherms of biochemical reactions can be considered, as illustrated in Fig. 2.5. The stability of such local equilibria is determined by the energy barrier between them. If this barrier is so low that thermal fluctuations can lead to quick transitions, the state is called *metastable*. This for example is the typical situation for enzyme-substrate complexes. For the schemes in Fig. 3.6 it must be considered that in reality G is not simply a function of a single reaction coordinate (x), but rather a hyperplane in n-dimensional space.

In contrast to equilibrium states, the stability conditions of the steady states are not simply dictated by their energy functions, but by their kind of entropy production. For this we must learn something more about this parameter.

In concordance with its definition the entropy production has the measure: $\mathrm{J\ K^{-1}\ s^{-1}}$. In some cases additionally this parameter is related to the mass in kg or to the molar mass.

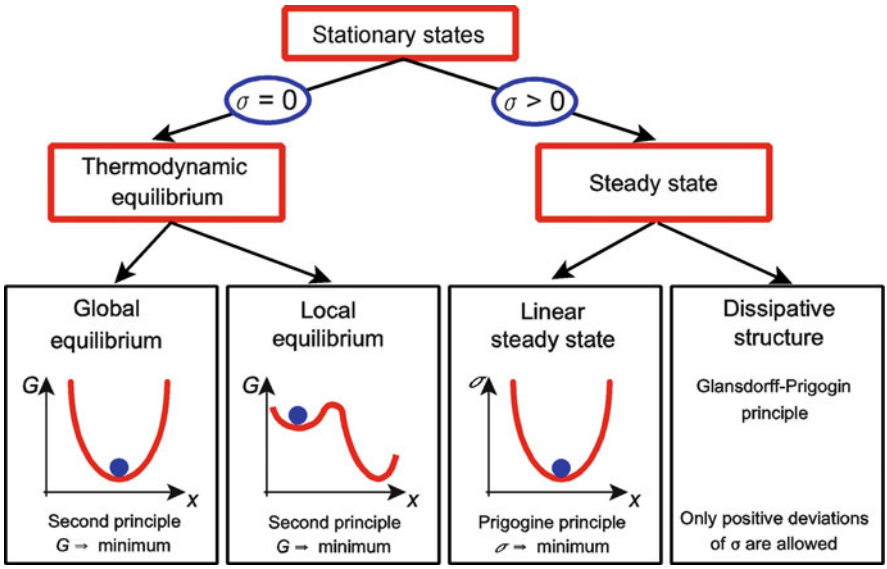

Fig. 3.6 Possible types of stationary states and conditions of their stability

Multiplying entropy production by temperature, one gets *Rayleigh's dissipation function* (**Φ**), which can be calculated from fluxes and forces as follows:

$$\Phi = \sigma T = \sum_{i=1}^{m} \mathbf{J}_i \mathbf{X}_i \tag{3.64}$$

This equation holds for the region of linear, as well as of nonlinear flux-force relations. Particularly for the linear region, Ilya Prigogine was able to show that systems tend to develop towards a reduced entropy production. This is the *Prigogine principle of minimal entropy production*. Systems, which are not far from thermodynamic equilibrium, and which are kept in imbalance by continuously acting forces consequently may move towards a steady state, the stability of which is included in this criterion (see Fig. 3.6).

The dissipation function represents the specific heat generation of a system. In living organisms it reflects the metabolic rate which in the case of aerobic organisms is equivalent to the oxygen consumption. Therefore, it can be measured directly by calorimetry or even calculated from parameters of respiration (see Sect. 3.8).

Some authors discuss this metabolic rate in context with this Prigogine principle of minimal entropy production. Figure 3.7 shows for example a mean scheme of the characteristic situation in mice. While the *total* heat production ($\dot{Q}$) increases in accordance with their age and mass (m), the *specific* heat production, that is the heat production relative to the body weight, i.e., the dissipation function (**Φ**), reaches a maximum and then decreases. It seems therefore that an animal in the state of

development increases entropy production, whereas an adult organism tends to arrive at a minimum. Disturbances in life lead to various deflections from this curve. If an organism for example is injured, if it is stressed by some environmental conditions, or in the case of tumor growth, the corresponding disturbances of the metabolism lead to a temporary increase in the slope of the curve.

It is still controversial, however, as to whether the Prigogine principle can be applied to large systems that include a great number of different subsystems, particularly those which are far from equilibrium. If a system deviates from the region of linear approaches, then the Prigogine principle is no longer valid. In contrast to steady states in the scope of linear thermodynamic approaches which are always stable and do not show any kind of metastability, systems in the region of nonlinear approaches show more complicated behavior.

Considering nonlinear systems we must first generalize the term "stationary." We already mentioned in context with Stoke's law (Eq. 3.54) that the so-called stationary movement is a movement with constant velocity where the frictional force is equal to the driving force. This sort of movement is also a kind of stationary, i.e., time-independent state. The term "stationary state" can also be applied to states that are not at rest, but show repetitive periodic movements in a stationary manner. For example, cardiac function can be called "stationary," if the frequency and amplitude of the heart beat does not change during the period of observation.

In the case of a linear steady state, fluctuations of the system parameters produce only positive deviations of entropy production, bringing the system back to the stationary state, which is therefore stable in any case. In contrast to this, in the region of nonlinear approaches, far from equilibrium, fluctuations of a stationary state can lead also to negative deviations of entropy production, leading to a destabilization of the system. The system may jump from one into another

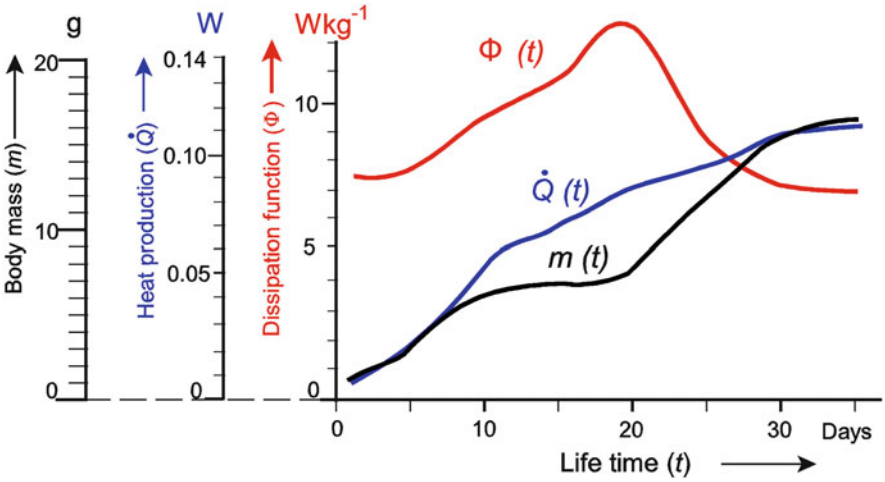

Fig. 3.7 Dissipation function (**Φ**), heat production ($\dot{Q}$), and mass (m) of mice as a function of their life span from the time of birth, up to the 35th day of life (Data from Prat and Roberge 1960)

stationary state. The steady states in the region of nonlinear approaches therefore are mostly metastable (see Fig. 3.5). Their stability condition requires the occurrence of only positive deviation of entropy production. This is the so-called *Glansdorff–Prigogine principle*.

The transition from linear, to nonlinear approaches of thermodynamics is not only associated with an increase in the coefficient of friction, or with the occurrence of several stationary states, but also with the spontaneous development of so-called *dissipative structures*. The general concept of the diversity of structures has already been discussed in Sect. 2.1.3. We postulated that there are basically, two different types of structures: equilibrium structures ($\sigma = 0$), and dissipative structures ($\sigma > 0$). In contrast to equilibrium structures which are always structures in space, dissipative structures can also be structures in time, or in a space–time continuum.

In order to illustrate this, Fig. 3.3 shows the function of the frictional coefficient $f(\mathbf{v})$ of a sphere, and the pattern of flow close to it. At a low velocity a laminar flow around the sphere occurs, changing into a turbulent flow, when the nonlinear region of this function is attained. When the flux-force relation changes from the laminar to the nonlaminar region, then the unstructured laminar flow becomes unstable and vortices appear which in the terminology of thermodynamics, are dissipative structures (for details of streaming behavior, see Sect. 3.7.1).

There exists an extensive literature on the theory of dissipative structures and on their occurrence in nature. Most of these dissipative structures are periodic structures in space, such as cloud patterns, flow phenomena in liquids with an applied temperature gradient, so-called Benard cells, plasma waves in electron beam tubes, etc. In addition, there are many time structures, including for example all kinds of sound production, from the electronic organ to vibrating strings, and to wind instruments.

In biological systems, in spite of many speculations dissipative structures in space have not been unequivocally demonstrated. Conversely, time patterns showing this property have been found to occur quite often. Examples of this include oscillations in metabolism, periodic changes in the membrane potential of certain cells, for example in the cells of the cardiac pacemaker in the sino-auricular node, and finally more complex, oscillatory movements in ecosystems. Sometimes, such oscillations in local concentrations of certain substances become "frozen," so that structures arise which do not require entropy-producing processes to sustain them, but which originally had been built as dissipative structures. One can conceive, for example, the genesis of the first viable biomacromolecule occurring in this way. Similar processes could form the basis of morphogenesis (for further detail, see Sects. 5.2.3, 5.2.4, and 5.2.5).

Let us come now to a further aspect of stationary states. If a system is considered as stationary, this is always in relation to a defined period of time. An organism could be said to be in a stationary state for a number of hours, or for days. That is, its volume, the content of certain substances, its shape, temperature, etc., are constant within defined tolerance for this period of time. Nevertheless, the organism in fact is ageing or growing. If a longer time period is considered, then the changes in these parameters will exceed the defined tolerance limits.

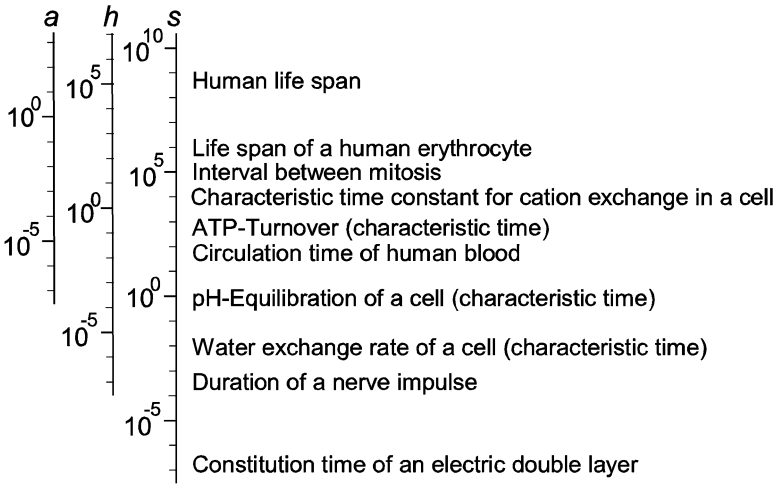

Fig. 3.8 Characteristic time constants (*s* – seconds, *h* – hours, *a* – years) of various processes on a logarithmic scale, to illustrate the time hierarchy in biological systems

A biological system consists of subsystems, each of which is associated with a different time constant. For example, the rate of aging in man is slow, in comparison with the rate of mitosis of a hemopoetic cell. Thus, the conditions in bone marrow can be regarded as stationary during the course of several cycles of mitosis in spite of the ageing processes affecting the organism as a whole. In vivo, the life span of a human erythrocyte is about 100 days. The ionic composition of juvenile erythrocytes differs somewhat from that of mature cells. If one is interested in the ionic regulation of these cells, then, because of the temporal characteristics of such processes, experiments lasting a few hours will be sufficient to provide the required information. Within such short time periods. the ionic concentration can be regarded as stationary. Figure 3.8 shows some characteristic times on a logarithmic scale that will serve to extend this list of examples.

These considerations hint at the existence of a *time hierarchy* of stationary states, related to their time constants which range over several orders of magnitude. The kinetic consequences of this circumstance will be considered in Sect. 3.2.5. The following concept, however, is important for our further thermodynamic considerations:

The living organism as a whole when considered within a limited period of time, is in a stationary state with entropy production, i.e., in a steady state. It is made up of a great number of subsystems which are ordered in a defined time hierarchy. The steady state of the system as a whole does not imply that all of the subsystems are also in a steady state. A large proportion of them, particularly those with a short time constant, are in thermodynamic equilibrium. If the system as a whole changes its parameters slowly, then these subsystems are capable of following such changes quickly, so that they almost completely adapt within their characteristic time, and thus are always in a stationary state. This is sometimes called a *quasi-stationary*, or *quasi-equilibrium* state.

The following example will illustrate this: The water content of a tissue depends on the ionic composition of its cells. Sodium and potassium ions are being actively transported against passive fluxes, giving rise to a steady state. In this way the active transport and the corresponding passive fluxes regulate the osmotic properties of the cells. The characteristic time of the water flux is much shorter than that of the cations (see Fig. 3.8). As a result, the water in the interior of the cells is always in osmotic equilibrium with the surrounding medium. In Sect. 3.2.3 we will discuss this example in detail in context with the Donnan-osmotic quasi-equilibrium.

Further Reading

Feistel and Ebeling 2011; Glansdorff and Prigogine 1985; Haken 2010; Zotin 1990.

3.1.5 Stochastic Resonance and Noise-Enhanced Processes

Noise-enhanced processes have been widely observed throughout nature. It is now broadly applied to describe any phenomenon where the presence of noise is better for output signal quality than its absence. It spans the field from diverse climate models, social and financial problems, via technical systems like electronic circuits, SQUIDs and lasers, to various biological systems such as ecological and neural models, ion channels, networks of biochemical reactions, etc.

Stochastic resonance (SR) as a particular property of dynamic systems is one of the mechanisms that makes a nonlinearity less detrimental to the noise of a signal. Quite unexpectedly, in this case noise can even lead to the formation of more regular temporal and spatial structures, and cause recognition and amplification of weak signals accompanied by growth of their signal-to-noise ratio (SNR). In fact, noise in this case can play a constructive or beneficial role in nonlinear dynamics.

The term stochastic resonance was first used in the context of noise-enhanced signal processing in 1980 to explain the almost periodic recurrence of the primary cycle of ice ages every 1,00,000 years. The application of SR in biology started in the early 1990s wherein SR was discovered in sensory neurons that had been subjected to external noise.

SR can be considered as a phenomenon occurring in nonlinear systems far from equilibrium, where the presence of a particular quantity of noise, superimposing a signal input is better for its output quality than its absence. The word “resonance” in this term on one hand was used because the plot of output performance, resp. the signal-to-noise ratio – against the noise intensity resembles a bell-shaped function with a single maximum; a similar appearance to frequency-dependent systems for some resonant frequency. On the other hand a kind of periodic resonance between properties of the system and some noise inherent frequencies occurs.

The basic mechanism of SR is schematically illustrated in Fig. 3.9. This example stands for a kind of excitation model of a sensory system. An external signal (red line) does not attain the threshold value (green line), therefore has no influence on the system. If however, the superimposed noise (blue) becomes large enough to arrive at the threshold value in cases of the maxima of the signal intensity, the

periodicity of the signal will be reflected in the system (black bars). In this case the noise becomes helpful for recognition of the signal by the system. If, however, the noise intensity increases further, the character of the signal will be fully masked and the SNR again tends to zero. This bell-shaped function of SNR in dependence of the noise intensity is depicted in the curve below.

This particular system of SR in the mechanism of nerve excitation is investigated quantitatively using the basic equations of the Hodgkin–Huxley model (Eqs. 3.195–3.200 in Sect. 3.4.4). It helps to understand the role of SR in many animal sensory systems.

To explain mechanisms of stochastic resonance in chemical reactions and various transport processes, the double-well potential profile of potential energy as a function of the reaction coordinate must be mentioned (cf. Fig. 2.5). In Sect. 2.1.5 we considered this system to deviate the Arrhenius equation. We mentioned there that even processes of diffusion and membrane transport in principle are based on the same kind of multiwell potential profiles.

In contrast to the equilibrium considerations of Sect. 2.1.5, let us now focus on the transition process between the minima of this function, including fluctuations induced by internal and external noise. For illustration one can apply a more mechanistic view, considering the state of a system like a position of a particle in a mechanical well (Fig. 3.10). The motion of such a particle qualitatively can be characterized by two time scales: The first defines relaxations of fluctuations in the linear regime around the stable fixed points (*intrawell dynamics*), the second concerns the mean time of barrier crossings (*global dynamics*), as a result of a nonlinear process.

The transformation of an external signal into an equivalent system behavior can be considered as a periodic transition from one state to another, or in a mechanistic view, as the movement of the particle from one well into the neighboring well. This is only possible in a reasonable time if the external signal transforms the energy function in such a way that the barrier between the two wells becomes small enough. Otherwise (Fig. 3.10a), this transition does not occur, in other words, the external signal will not be reflected by system behavior. In the case of an additional noise (Fig. 3.10b), however, this becomes possible. In the same way as depicted in Fig. 3.9, the external signal at optimal noise intensity can be received by the system.

In this context an interesting case of biological optimization should be noted. If a particular noise intensity in some sensory systems is necessary to achieve optimal SR, and if this noise has its origin in the stochastic properties of ion channels, optimization can occur as a cooperation of several channels. The noise of a single channel can be reduced by averaging the effect of an assembly of other channels. Conversely, the SR effect vanishes if the assembly becomes too large. Apparently, an optimal size of cooperating channel assemblies exists in such systems.

Thus, stochastic resonance allows us to realize the input of a signal at an optimal noise level by switching events in a bistable system. Similarly, noise can enhance oscillations in a dynamic system even without external periodic signals. This phenomenon is called *coherence resonance*.

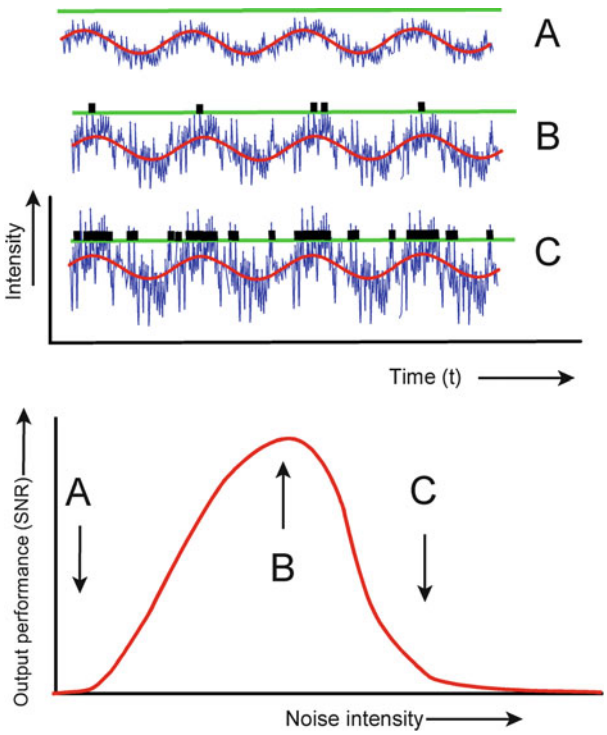

Fig. 3.9 A scheme explaining the mechanism of stochastic resonance in a system with a definite threshold like a neuron. *Above* – A, B, C: a subliminal sinusoidal signal (*red line*), superimposed by noise (*blue*) with increasing intensity. A – weak noise, not attaining the threshold intensity (*green line*). B – optimal noise intensity, leading to excitations (*black bars*), the time course of which reflect the periodic character of the signal. C – extensive noise intensity, smearing the character of the signal. *Below*: The output performance as a bell-shaped function of the noise intensity

This situation is modeled for example in nerve cells using a simplified version of the Hodgkin–Huxley equations, which accounts for the essentials of regenerative firing mechanisms. In fact, neurons are permanently affected by various kinds of noise, such as for example the fluctuating opening and closing of ion channels, the noise of presynaptic currents, fluctuations of conductivities, etc. As the neuron is unable to fire during the recovery state, it becomes excited at a particular time scale by sufficient intense noise. In this way an oscillatory behavior occurs. Since the excitation time depends strongly on the noise, an optimal noise level exists like in SR systems. However, unlike stochastic resonance, in the case of coherence resonance no particular input signal exists.

There are also several other mechanisms, distinct from SR, where noise has a constructive role. An example is the so-called *Brownian ratchet*, a mechanism, which is used in technics, for example in some battery-less wristwatches that are wound up by random movement. In this case a stochastically oriented force

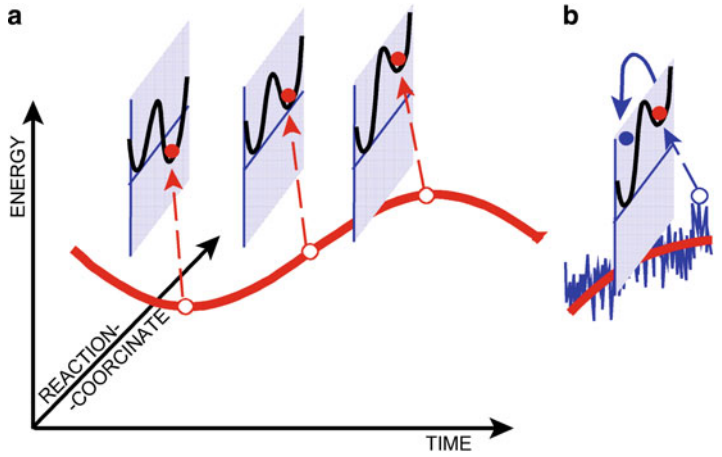

Fig. 3.10 Stochastic resonance in terms of the reaction coordinate. (**a**) A periodic signal (*red line*) modifies periodically one part (the backmost) of the double well potential of a reaction, but not sufficiently to lead the system (illustrated as the position of the *red particle*) in a reasonable time to transit into the neighboring (in front) well (A). (**b**) This however occurs under the influence of a sufficiently intense noise. The effective transition of the particle between these two states in this case may reflect the time scale of the periodic signal (*red line*) in the same way as in the example of Fig. 3.9

generates a directed movement. The physical basis of this phenomenon is an oscillating force, affecting an anisotropic system, where friction in one direction differs from that in the opposite.

This mechanism led to speculations that such a Brownian ratchet, like a Maxwell demon could be driven by thermal fluctuations, and in this way would form a perpetuum mobile of the second order (see Sect. 2.1.2). A Brownian particle was considered to interact with a ratchet exhibiting a sawtooth profile. In this case it could move forward, driven simply by thermal noise. This would represent a microscopic version of the winding mechanisms that are realized in wristwatches. In accordance with the second law of thermodynamics, however, it is impossible to produce oriented movement driven by thermal oscillations. Thus, as long as the pawl has the same temperature as the ratchet it is subjected to the thermal noise itself, i.e., its particular structure will be disturbed in molecular dimensions. In fact the self-winding mechanism of the wristwatch is driven not by thermal noise but by stochastically oriented acceleration, the energy consuming movement of the arm. Its mechanism is based not on a rectifying of thermal noise, but on an input of energy consuming stochastic force.

In biological systems various types of Brownian ratchets, as mechanisms rectifying stochastically oriented forces, are identified, on the molecular level as well as on the level of movement of cells and organisms. So for example, there exist complex patterns of directed molecular movement that are based on filaments and motor molecules, performing mechanical work on the nanometer scale driven by ATP hydrolysis. These molecular motors and motor particles bind to cytoskeletal

filaments and walk along these filaments in a directed fashion. This can be understood in terms of Brownian ratchets driven by nonequilibrium processes. The motor molecules can be considered as Brownian particles which can attain several internal states, and experiences a certain molecular force potential in relation to the cytoskeletal filament.

Further Reading

Anishchenko et al. 2006; Hänggi 2000; Lipowsky and Klumpp 2005; Mahmud et al. 2009; McDonnell and Abbott 2009.

3.1.6 Thermodynamic Basis of Biochemical Reactions

Chemical processes in biological systems are very complex and specific. In Sect. 2.1.5, we already mentioned the ability of enzymes to overcome energy maxima of activation energy. Some aspects of chemical thermodynamics are explained in Sect. 3.2.1. In Sect. 5.2.1, we will discuss the particularities of the kinetics of biochemical reactions. Now, some general thermodynamic aspects of chemical and biochemical reactions will be introduced to complete the approaches of equilibrium as well as nonequilibrium thermodynamics.

In the previous formulation of the Gibbs equation (Eqs. 3.17, 3.27, 3.28, 3.29), the chemical reaction was not explicitly enclosed. In fact, a chemical reaction can be analyzed by referring to the chemical potential of its components. The reaction:

$$\nu_a \mathrm{A} + \nu_b \mathrm{B} \rightarrow \nu_c \mathrm{C} + \nu_d \mathrm{D} \tag{3.65}$$

can be considered as a replacement of the substances A and B by the substances C and D.

Considering this process for isobaric ($\mathrm{d}p = 0$) and isothermic ($\mathrm{d}T = 0$) conditions, where only concentration changes take place ($\mathrm{d}n_i \neq 0$, $\mathrm{d}q = 0$, $\mathrm{d}l = 0$), the equation for Gibbs free energy (Eq. 3.29) reduces to:

$$\mathrm{dG} = \sum_{\mathrm{i}=1}^{\mathrm{m}} \mu_{\mathrm{i}} \ \mathrm{dn_i} \tag{3.66}$$

The change in the Gibbs free energy ($\Delta_R G$) according to the definition (3.31) is then given by:

$$\Delta_R G = \nu_c \mu_c + \nu_d \mu_d - \nu_a \mu_a - \nu_b \mu_b \tag{3.67}$$

Substituting in this equation the expression for the chemical potential (Eq. 3.33), we obtain:

$$\begin{aligned}\Delta_R G^0 = {} & \nu_c \mu_c^0 + \nu_d \mu_d^0 - \nu_a \mu_a^0 - \nu_b \mu_b^0 + \\ & + \mathrm{R}T(\nu_c \ln a_c + \nu_d \ln a_d - \nu_a \ln a_a - \nu_b \ln a_b)\end{aligned} \tag{3.68}$$

Now, we can summarize these standard potential expressions and define a molar free standard reaction energy:

$$\Delta_R G^0 = \nu_c \mu_c^0 + \nu_d \mu_d^0 - \nu_a \mu_a^0 - \nu_b \mu_b^0 \tag{3.69}$$

Substituting this in Eq. 3.68 and combining the logarithmic terms, we obtain the *van't Hoff equation*:

$$\Delta_R G = \Delta_R G^0 + \mathrm{R}T \ln \frac{a_c^{\nu_c} a_d^{\nu_d}}{a_a^{\nu_a} a_b^{\nu_b}} \tag{3.70}$$

In the case of thermodynamic equilibrium, $\Delta_R G = 0$. This allows to define the equilibrium constant (K_p) for reactions under isobaric conditions, whereas the symbols a_i^0 stand for activities of the system in equilibrium:

$$\mathrm{K}_p \equiv \frac{a_c^{0\nu_c} a_d^{0\nu_d}}{a_a^{0\nu_a} a_b^{0\nu_b}} = e^{-\frac{\Delta_R G^0}{\mathrm{R}T}} \tag{3.71}$$

Substituting this expression into Eq. 3.70, we obtain for the case of equilibrium ($\Delta_R G = 0$):

$$\Delta_R G^0 = -\mathrm{R}T \ln \mathrm{K}_p \tag{3.72}$$

The molar free standard Gibbs energy of a chemical reaction ($\Delta_R G^0$) can be calculated using the standard energies of formation ($\Delta_F G^0$) obtained from tables (see Eq. 3.31). In fact, we use here the property of Gibbs energy as being a state function, that is, a parameter which is independent of the way in which it is obtained (see Sect. 3.1.1).

If a reaction is not in equilibrium, then the direction which it will take can be determined by calculating $\Delta_R G$ (Eq. 3.70) from the given activities of the components. Spontaneously the reaction can run in the direction indicated in Eq. 3.65 only in the case where $\Delta_R G < 0$, resulting in a reduction of the free energy during this process.

In this approach, we considered chemical reactions simply by changing the concentrations of their compounds and not considering whether the alterations of these concentrations were realized by transport processes or actually by a chemical reaction. This is, in fact, more appropriate and useful in most cases. In some cases,

however, chemical reactions take place combined with the processes of transport of matter through biological membranes. In such cases, this approach can be misleading. In this case it is useful to introduce a special term into the Gibbs equation differentiating between transport processes and the chemical reaction.

In fact, it is possible to measure the progress of a chemical reaction using the definition of the *degree of advancement* ($d\zeta$):

$$d\xi = \frac{1}{\nu_i} dn_i \tag{3.73}$$

A positive value of $d\zeta$ represents a particular step in the reaction from left to right, whereas $dn_i > 0$ indicates an increase and $dn_i < 0$ a decrease of the molar concentration of the substance. In order to maintain these sign conventions, the stoichiometric coefficients (ν_i) of the initial substrates must become negative and those of the end products, in contrast, positive.

Taking, for example, the reaction

$$2H_2 + O_2 \Rightarrow 2H_2O$$

the following applies:

$$d\xi = -\frac{1}{2} dn_{H_2} = -dn_{O_2} = \frac{1}{2} dn_{H_2O} \tag{3.74}$$

Now, as the work coefficient according to our assumption in Sect. 3.1.2, the *chemical affinity* (A) is defined as:

$$A = -\sum_{i=1}^{m} \nu_i \mu_i \tag{3.75}$$

where $i = 1 \ldots m$ are the components of the given reaction. Using Eq. 3.73, the following relation is obtained:

$$\sum_{i=1}^{m} \mu_i \; dn_i = -A \, d\xi \tag{3.76}$$

To differentiate real chemical reactions ($d\zeta$) and transport processes (dn_i), it is useful to substitute both expressions into the Gibbs equation simultaneously. This may be important when determining the processes of active ion transport through membranes.

Affinity, from its definition, proves to be an energy difference and, as such, the driving force of a chemical reaction. Consequently, it can be directly inserted into the flux matrix (Eq. 3.58) instead of a generalized force (**X**).

3.2 The Aqueous and Ionic Equilibrium of the Living Cell

From the point of view of thermodynamics, a living cell can be considered in general as a system in nonequilibrium. This nonequilibrium state is maintained by permanent processes of energy transformation. In spite of the state of the whole system, however, some of these subsystems nevertheless may be in thermodynamic equilibrium. The occurrence of such equilibria, or quasi-equilibria have already been discussed in Sect. 3.1.4, in connection with the time hierarchy of biological systems. In this section we will direct our attention to these equilibrium states, especially equilibrium distributions of charged and uncharged components.

3.2.1 The Van't Hoff Equation for Osmotic Pressure

It is very important to understand the concept of osmotic pressure for a number of cell physiological processes. Unfortunately, osmotic pressure is often confused with hydrostatic pressure, especially with the turgor pressure in plant cells. In this section it will be shown that osmotic pressure is just a property of a solution or a suspension and that although it can be the origin of the hydrostatic pressure in a cell, it is not at all identical to it.

The effect of osmotic pressure can be demonstrated best by a *Pfeffer cell* (see Fig. 3.22). This consists of a glass bell with a vertical tube or a manometer at the top. The mouth of the bell is closed by a semipermeable membrane. The bell is filled with a solution and submerged in a vessel with pure water. The membrane allows the water, but not the molecules of the solute, to pass through. In his experiments in 1877, the botanist Wilhelm Pfeffer used a sheet of porous pottery to close the bell. The pores of this "membrane" were covered by a precipitated layer of cupric(II)-hexacyanoferrate(II). In this experiment the water penetrates the membrane, passes into the glass bell, and increases the internal hydrostatic pressure. Eventually an equilibrium will be reached when the internal pressure balances the osmotic force which drives the water into the Pfeffer cell. If the membrane is truly semipermeable, if pure solvent is outside the glass bell, and if the system is really in equilibrium, then, and only then, does the hydrostatic pressure difference between the internal and the external phase in equilibrium state equal the osmotic pressure of the solution inside the Pfeffer cell.

In order to analyze this situation let us assume two phases (I and II) separated from each other by a membrane (see Fig. 3.11). Let the membrane be semipermeable thus allowing the solvent (w), but not the solute (s), to pass through. Because of solvent flux through the membrane, the volumes of both phases and their concentrations will change. This will also alter the chemical potentials (μ_i) of the components (i).

The concentration dependence of the chemical potential, using the mole fraction (x_i) (see Eq. 3.33) as a measure of concentration is:

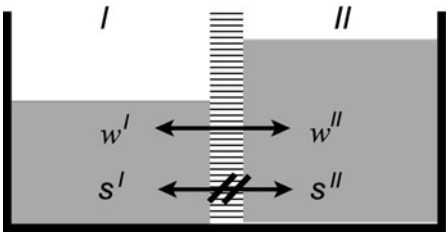

Fig. 3.11 Derivation of osmotic pressure: Phases I and II are separated from each other by a semipermeable membrane. Only the solvent (w), but not the solute (s), can penetrate this membrane. The hydrostatic pressure of both sides of the membrane is different

$$\mu_i = \mu_{ix}^0 + \mathrm{R}T \ln x_i \tag{3.77}$$

To achieve thermodynamic equilibrium under isothermal conditions, the chemical potential of the exchangeable components of both phases I and II must be equal. In the example considered here, this will only apply to the solvent (w). This requires

$$\mu_w^I \stackrel{!}{=} \mu_w^{II} \tag{3.78}$$

which means that

$$\mu_{wx}^{0I} + \mathrm{R}T \ln x_w^I = \mu_{wx}^{0II} + \mathrm{R}T \ln x_w^{II} \tag{3.79}$$

or

$$\mathrm{R}T \ln \frac{x_w^I}{x_w^{II}} = \mu_{wx}^{0II} - \mu_{wx}^{0I} \tag{3.80}$$

Now, let us first direct our attention to the difference in the standard chemical potentials on the right-hand-side of Eq. 3.80. Because of the pressure difference induced by the flow of the solvent, $\mu_{wx}^{0I} \neq \mu_{wx}^{0II}$. The pressure dependence of the standard chemical potentials therefore must be studied in more detail.

The following definition of the standard chemical potential can be derived from Eq. 3.32:

$$\mu_i^0 = \left(\frac{\partial G^0}{\partial n_i}\right)_{T,p,n_j} \tag{3.81}$$

The pressure dependence of μ_i^0 can therefore be expressed as

$$\frac{\partial \mu_i^0}{\partial p} = \frac{\partial \left(\frac{\partial G^0}{\partial n_i}\right)}{\partial p} = \frac{\partial \left(\frac{\partial G^0}{\partial p}\right)}{\partial n_i} \tag{3.82}$$

The way in which the sequence of the derivation steps in a total differential can be changed was discussed in Sect. 3.1.1 (see also Eq. 3.6).

From the equation of Gibb's free energy (Eq. 3.29) we obtain:

$$\mathrm{d}G^0 = V\ \mathrm{d}p \quad \text{for}: \quad \mathrm{d}T = 0, \quad \mathrm{d}l = 0, \quad \mathrm{d}q = 0, \quad \mathrm{d}n = 0 \tag{3.83}$$

and

$$\frac{\partial G_i^0}{\partial p} = V_i \tag{3.84}$$

If this is substituted into Eq. 3.82, then

$$\frac{\partial \mu_i^0}{\partial p} = \frac{\partial V_i}{\partial n_i} = \bar{V}_i \tag{3.85}$$

We are already acquainted with the variable $\bar{V}$ which is the partial molar volume of the substance i (Eq. 3.8). Now, the differential $\mathrm{d}\mu_i^0$ can be calculated. This is a small change in μ_i^0 that occurs when there is a small change in the pressure $\mathrm{d}p$. Following the general rule (3.1), we obtain:

$$\mathrm{d}\mu_i^0 = \left(\frac{\partial \mu_i^0}{\partial p}\right) \mathrm{d}p = \bar{V}_i\ \mathrm{d}p \tag{3.86}$$

To obtain an expression for the difference in the standard potentials due to alterations in pressure, the equation must be integrated between the corresponding limits:

$$\int_{\mu_i^{0I}}^{\mu_i^{0II}} \mathrm{d}\mu_i^0 = \int_{p^I}^{p^{II}} \bar{V}_i\ \mathrm{d}p \tag{3.87}$$

This leads to:

$$\mu_i^{0II} - \mu_i^{0I} = \bar{V}_i(p^{II} - p^I) \tag{3.88}$$

(This manner of integration only applies when $\bar{V} = const$, that is, when the partial molar volume is independent of the pressure. This is only the case for ideal solutions.)

Now, it is possible to substitute Eq. 3.88, related to the solvent (w) into Eq. 3.80:

$$p^{II} - p^I = \frac{\mathrm{R}T}{\bar{V}_w} \ln \frac{x_w^I}{x_w^{II}} \tag{3.89}$$

This equation indicates the hydrostatic pressure difference ($p^{\mathrm{II}} - p^{\mathrm{I}}$) induced in our system at thermodynamic equilibrium for different values of the mole fractions in the two phases $x_w^I \neq x_w^{II}$. If phase I contains only pure solvent ($x_w^I = 1$), this pressure difference is called osmotic pressure (π), therefore

$$\pi = \frac{\mathrm{R}T}{\bar{V}_w} \ln \frac{1}{x_w^{II}} = -\frac{\mathrm{R}T}{\bar{V}_w} \ln x_w^{II} \tag{3.90}$$

This equation and its derivation allow us to describe osmotic pressure as a parameter reflecting a property of a solution. Under isothermal conditions the osmotic pressure of a solution is equal to the hydrostatic pressure, which is required to alter the chemical potential of the pure solvent in such a way that it will be equal to the chemical potential of the solvent in this solution.

Equation 3.90 is a precise expression to calculate the osmotic pressure of a solution, provided the mole fraction activity of the solvent x_w^{II} is used to express the concentration of the solvent (w) in the solution. This is indeed a quite unusual and rather cumbersome form for the equation. How can this equation be transformed into a more simplified form?

Some simplifications can be made for dilute solutions. First, the *activity*, which as a matter of fact is what x_W in Eq. 2.14 means, can be replaced by mole fraction *concentration*. Using Eq. 3.35, this means:

$$x_w = \frac{n_w}{n_s + n_w} \tag{3.91}$$

If the sum of the mole fractions of all components of a solution equals 1 (Eq. 3.36), then:

$$x_w = 1 - x_s \tag{3.92}$$

The number of molecules of solvent in diluted solutions is much larger than the number of molecules of the solute ($n_w \gg n_s$). A 0.1-molar solution of the substance (s) in water (w), for example, contains $n_s = 0.1$ mol of solute, but $n_w = 55.6$ mol of water (n_w means moles of water per 1,000 g solution, therefore, 1,000/18 = 55.6). In this case, using the definition of x_S (Eq. 3.35) it becomes:

$$x_w = 1 - \frac{n_s}{n_s + n_w} \approx 1 - \frac{n_s}{n_w} \tag{3.93}$$

In addition, the following holds for dilute solutions:

$$\bar{V}_w = \frac{\partial V_w}{\partial n_w} \approx \frac{V_w}{n_w} \tag{3.94}$$

Thus, $n_w = V_w/\bar{V}$. One can also introduce the molar concentration of the solute (s), using $n_s/\mathrm{V} = c_s$, and the total volume of the solution ($V \approx V_w$). Substituting this in Eq. 3.93, we obtain:

$$x_w = 1 - \frac{n_s \bar{V}_w}{V} = 1 - c_s \bar{V}_w \tag{3.95}$$

Let us substitute this expression in Eq. 3.90:

$$\pi = -\frac{\mathrm{R}T}{\bar{V}_w} \ln(1 - c_s^{II} \bar{V}_w) \tag{3.96}$$

Now, we use the following rule to expand logarithmic expressions in series. For any number $|q| < 1$, it holds:

$$\ln(1 - q) = -q - \frac{q^2}{2} - \frac{q^3}{3} - \frac{q^4}{4} - \ldots \tag{3.97}$$

For q we use the expression $c_s^{II}\, \bar{V}_w$, which is much smaller than 1. In this case we are justified in retaining only the first term of this series. This gives

$$\pi = -\frac{RT}{\bar{V}_w}(-c_s^{II} \bar{V}_w) \tag{3.98}$$

and

$$\pi = \mathrm{R}Tc_s^{II} \tag{3.99}$$

This is the *Van't Hoff equation* for osmotic pressure.

It is interesting to note that, in 1877, Van't Hoff derived this relation not in the way shown here, but by considering the analogy with the state equation for an ideal gas. He assumed that 1 mol of an ideal gas, when confined to a volume of 1 L, would exert a pressure of 2.27 MPa on the walls of the vessel. He reasoned that the molecules of a solute, when dissolved at a concentration of 1 mol/l should behave in the same way as the particles of the gas. This pressure he called "osmotic."

The equation of state for an ideal gas is:

$$p = \frac{n}{V}\mathrm{R}T \tag{3.100}$$

This can be transformed into:

$$\pi = \frac{n}{V}\mathrm{R}T = c\mathrm{R}T \tag{3.101}$$

Both the thermodynamic derivation as well as the analogy used by Van't Hoff show that Eq. 3.99 is only applicable for ideal solutions, or with some approximation for solutions that are very diluted. This restriction can be overcome by using a correction factor. This is the so-called *osmotic coefficient* (g):

$$\pi = gc\mathrm{R}T \tag{3.102}$$

This factor must not be confused with the activity coefficient (f), which was introduced in Sect. 3.1.2 (Eq. 3.34). This will be easily understood if f is directly introduced into the thermodynamic derivation of the Van't Hoff equation (e.g. in Eq. 3.79). In this case, it would reflect the activity coefficient of water, not that of the solute. The successive transformation of this equation up to Eq. 3.25, introducing the concentration of the solute (c_S) instead of the mole fraction activity of water (x_w) in fact used other assumptions. The relation between f and g is somewhat complicated and will not be discussed here.

In addition to the nonideal behavior of a solution, it must be considered furthermore that the osmotic pressure exerted by dissociable substances in fact results from the sum of the osmotic pressures of all the ions formed by dissociation. Thus, if a mono-monovalent salt (such as NaCl) is assumed to be completely dissociated into Na^+ and Cl^- then the osmotically active concentration is twice as great as the salt concentration.

Therefore, one must distinguish between the *osmolarity* and the *molarity* of a solution. The osmolarity of a 0.1-M solution of NaCl (at 25°C), is therefore: $2 \times 0.1 \times \mathrm{g} = 2 \times 0.1 \times 0.932 = 0.1864$ osM. The osmolarity can be pH-dependent for polyelectrolyte solutions because it will change according to the degree of dissociation. In contrast to the *osmolarity*, expressed as osmol/l, sometimes the term *osmolality* is used as a measure expressed as osmol/kg. Considering the temperature dependence of the volume, osmolality, i.e., relation to the mass in osmometry is preferred in relation to osmolarity.

The osmolarity of a solution can be measured by various methods. The direct method is the *membrane osmometer* where the hydrostatic pressure is indicated, resulting in a solution separated from pure solvent by a semipermeable membrane, like in a Pfeffer cell (schematically shown in Fig. 3.22). In this equipment, however, sensitive pressure sensors are used instead of the vertical tube with a water column. This method is applicable if membranes are available that are actually semipermeable to the particular solutes.

Mostly, the osmotic pressure of a solution is measured in an indirect way. In fact, the reduction of the vapor pressure, or the freezing point of a solution are basically determined in the same way as their osmolarity. Therefore, instruments are used to measure these properties. In the case of clear solutions the most accurate and comfortable technique is *cryoscopy*, i.e., the measurement of the freezing point. For solutions of macromolecules, and for polydisperse suspensions, however, this method cannot be used. In this case *vapor pressure* osmometers are preferred. They indicate small stationary temperature differences between a drop of the fluid, and a

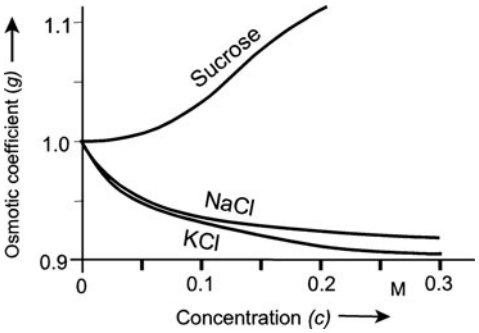

Fig. 3.12 The osmotic coefficient (g) as a function of the concentration (c) in aqueous solutions

drop of a control solution with known osmolality, caused by differences of evaporation in a definite atmosphere.

Relating the osmolarity of solutions given theoretically by Eq. 3.25, with the measured values, the osmotic coefficient (g) can be determined directly. Figure 3.12 indicates that for solutions of NaCl and KCl, starting from the value 1 in extremely dilute, i.e., "ideal" solutions, it drops to values near to 0.9 with increasing concentrations. The opposite behavior of sucrose is caused by its influence on the water structure. We will come back to this point, as well as to the relation between osmotic (g) and activity (f) coefficients in the next section.

3.2.2 Osmotic Pressure in Cells and Biologically Relevant Fluids

In the previous chapter basic equations were derived which help to understand osmotic pressure as a property of a solution, and allow exact (Eq. 3.90) – or with good approximation (van't Hoff Eq. 3.102) – calculation of this quantity. In order to apply the van't Hoff equation to nonideal solutions, a correction term was included, the osmotic coefficient (g). It usually becomes <1 if the concentration of the solute increases (see Fig. 3.12), in a similar way to the coefficient of activity (f) as shown in Fig. 3.1. The function $g(c)$, however, does not decline as strongly as the activity coefficient $f(c)$. But in both figures it is exceptional that the values for sucrose are rising. What could be the reason for this? What happens with other organic molecules, especially with the highly concentrated macromolecules in biological compartments? This question is important as the osmotic pressure is not only a property of true solutions, but also of colloidal solutions and to some extent also of suspensions. In this context it is referred to as *colloid-osmotic pressure*.

To solve this problem, the real properties of water in these solutions must be considered in detail. We already discussed the interaction of water dipoles with polar and apolar molecular compounds as well as with various kinds of surfaces (see Sect. 2.2.3). We asserted that several kinds of fixed, or even encapsulated water molecules in these kinds of solutions exist, which certainly contribute to the osmotic property to different degrees.

In fact, even in sucrose solutions, more so in suspensions, and in the cytoplasm of living cells, water exists both in osmotically active and in osmotically inactive phases, which may be even caused by these fixed, or "bound" kinds of water molecules. This is obvious, when analyzing the real osmotic behavior of these liquids.

To show this, let us introduce a molal *concentration* c_s', representing the true concentration of the solvent (s) in regard to the amount of osmotically active water. Therefore it is defined as:

$$c_s' = \frac{m_s}{m_{fw}} = \frac{m_s}{m_{tw} - m_{bw}} \tag{3.103}$$

where m are mol masses of s = solute, fw = "free" water, tw = "total water," bw = "bound" water. Now, we can define W as a measure of "bound" water as:

$$W = \frac{m_{bw}}{m_s} \tag{3.104}$$

Introducing m_{bw} from Eq. 3.104 into Eq. 3.103, rearranging, and replacing $m_s/m_{tw} = c_s$, results in:

$$c_s' = \frac{c_s}{1 - Wc_s} \tag{3.105}$$

The van't Hoff equation (Eq. 3.102) can now be used to introduce the "true" osmotic active concentration c_s', corrected with a "true" osmotic coefficient g'. For simplification we are using the osmotic pressure in the unit "osmolal," not in "Pa" as in Eq. 3.102. Therefore, we do not need the RT-term. It becomes:

$$\pi = g'c_s' \tag{3.106}$$

Introducing this into Eq. 3.105 and rearranging, results in:

$$\frac{1}{c_s} = W + \frac{g'}{\pi} \tag{3.107}$$

If now the value $1/c_s$ is plotted against $1/\pi$, a linear function appears, indicating W as the point at which the ordinate is crossed at $1/\pi \rightarrow 0$, and g' as the slope. This is done in Fig. 3.13 for sucrose, polyethylene glycole (M = 400), and bovine serum albumin (BSA) (to fit in the same figure, the molal and osmolal values for BSA were multiplied by 100!). The dotted line shows the slope for the case of $g' = 1$, which corresponds to all of the three lines. The intercepts (W) are obtained at 0.0956 for sucrose, 0.4094 for PEG400, and 189.48 for BSA (considering the factor of 100!). Converting this in terms of mol H_2O/mol solute, they are five for sucrose, 22.7 for PEG400, and 10,500 for BSA. Sucrose therefore is hydrated with one water

per OH-group, whereas BSA is surrounded by multilayers of water that do not participate in osmotic activity (according to Garlid 2000).

The osmotic pressure of the surrounding medium is an important parameter for the survival of cells, especially for their volume regulation. In classical physiology it is postulated that cells shrink in *hypertonic solutions* ($\pi_{sol} > \pi_{cell}$), attain their volume in *isotonic solutions* ($\pi_{sol} = \pi_{cell}$), and swell in *hypotonic solutions* ($\pi_{sol} < \pi_{cell}$). This is based on the assumption that cells behave like osmometers, the membrane of which is freely permeable to water but impermeable for all solutes. Furthermore, all the cellular water is considered as totally osmotically active.

There are two reasons why this assumption can be considered as strongly simplified. First, cells even more so than macromolecules in solutions contain water which is osmotically inactive. This can be demonstrated by considering the simplest animal cell: the mammalian erythrocyte, which can be considered as just a membrane-enclosed volume, packed with a large concentration of hemoglobin. Usually, the volume of these erythrocytes is measured as centrifuged hematocrit values, or by coulter counter measurements. Both methods may lead to artefacts, especially if solutions of different compositions are used. Therefore in Fig. 3.14 the relative volume is expressed as the inverse of protein concentration ($m_{water}/m_{protein}$ in g/g).

Similar to Fig. 3.13, the volume as an inverse of a concentration is plotted against the reciprocal osmolarity of the external medium. The resulting curve in the same way as in Fig. 3.13 does not attain the zero point of the ordinate, clearly indicating that erythrocytes contain osmotically inactive water. Furthermore, the measured points seem to show three different linear segments. In the region of isotonic osmolality, this reflects in fact the behavior of native hemoglobin with a

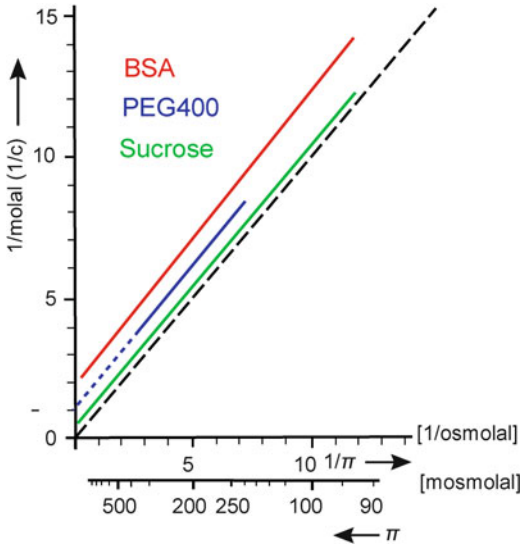

Fig. 3.13 The plot of measured osmotic activity of bovine serum albumin (BSA), polyethylene glycole (PEG400), and sucrose according to Eq. 3.107. Molal and osmolal values for BSA were multiplied by 100 to fit in the same figure (After Garlid 2000)

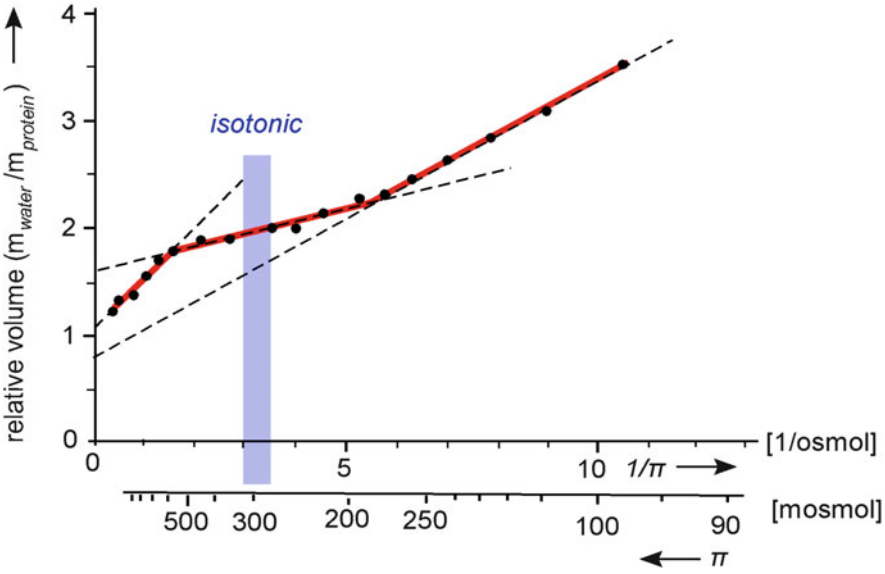

Fig. 3.14 Osmotic behavior of human erythrocytes in NaCl-solutions of various osmolarity. The relative volume is measured in mean water content ($m_{water}/m_{protein}$) (After Fullerton et al. 2006, redrawn)

content of fixed water of about 1.6 g per gram protein. Also shown are deviations away from isotonicity. In the hypotonic region, obviously a number of osmotically active small-sized molecules are released, and in the shrunken cells in hypotonic media hemoglobin aggregates. Similar experiments with corresponding results have been performed also in mitochondria.

The second, nonrealistic approach in the oversimplified consideration of the cell as a simple osmometer, concerns the question of semipermeability. Investigating for example erythrocytes in sucrose solutions, the permeability of water is in fact much greater than that of the external solutes. Using, however, smaller molecules in the external solution, like glucose, urea, etc., the cell membrane cannot be considered semipermeable and furthermore, the conditions used for the derivation of the osmotic equations do not hold. The cell will not arrive at a thermodynamic equilibrium as long as the solute is penetrating the membrane.

This question ultimately concerns the relation between the difference between osmotic ($\Delta\pi$) and hydrostatic pressure (Δp). Only in the case of thermodynamic equilibrium of water, and only if the membrane is actually semipermeable to all components of the solution, does the difference in osmotic pressure equal the difference in the generated hydrostatic pressure.

In the case of solutions with several components indicating various degrees of permeability, the following relation between osmotic ($\Delta\pi$) and hydrostatic (Δp) differences can be applied:

$$\Delta p = \sum_{i=1}^{n} \sigma_i \Delta \pi_i \tag{3.108}$$

This equation takes into account that in the system n substances, each with an osmotic difference $\Delta\pi_i = \pi_{i(\text{internal})} - \pi_{i(\text{external})}$, determine the osmotic conditions. Their effectiveness, with regard to the development of hydrostatic pressure, depends on the value of a factor (σ_i) which is known as *Staverman's reflection coefficient*. In contrast to the "classical" approach, this model takes into account that the membrane is not semipermeable, but permselective. This means that all components of the solution can more or less penetrate the membrane. We will analyze this situation in detail later using the approaches of nonequilibrium thermodynamics (Sect. 3.3.1). Upon consideration of the corresponding flux matrix (Eq. 3.147) the following relation is derived:

$$\sigma_s = \frac{\nu_w - \nu_s}{\nu_w} \tag{3.109}$$

Using the indices of the Van't Hoff's equation, ν_w and ν_s represent the rate of movement of the solvent (water) and the solute in the membrane. In the case of $\nu_s \to 0$, the reflection coefficient becomes $\sigma_s \to 1$. This is the "classical" situation of semipermeability, and therefore $\Delta p \to \Delta\pi$. However, when $\nu_s \to \nu_w$, then $\sigma_i \to 0$ and hence $\Delta p \to 0$. This occurs in the case of a membrane which allows the osmotically active substance, as well as the solvent to pass through. In this case, no hydrostatic pressure can develop even at initial differences in osmotic pressure. In such a system a thermodynamic equilibrium would finally lead to a homogeneous distribution of substance i (see Fig. 3.22).

In general, the reflection coefficient for disaccharides, such as sucrose, and for larger molecules equals nearly one. Smaller molecules, especially those which can directly penetrate the lipid layer of a biological membrane, show lower values (see Table 3.1).

Table 3.1 Typical values of reflection coefficients of nonelectrolytes for human erythrocytes (Giebisch et al. 1979) and for *Nitella flexilis* (Zimmermann and Steudle 1978). Values in parentheses: Levitt and Mlekoday 1983

	Erythrocytes	*Nitella*
Urea	0.79 (0.95)	0.91
Thiourea 0.91		
Ethylene glycol	0.86 (1.0)	0.94
Glycerol 0.88	0.80	
Acetamide 0.80	0.91	
Propionamide	0.84	
Malonamide	1.00	
Sucrose	0.97	
Glucose	0.96	
Methanol	0.31	
Ethanol	0.34	
Isopropanol		0.35
n-Propanol	0.17	

In plant cells, osmotic differences generate the so-called *turgor pressure*, which can be as high as several hundred kPa. This intracellular pressure plays a large role in the mechanical stabilization of plants. It forces the cell membrane of the plant cell against the mechanically stabilizing cell wall, which itself is freely permeable to ions and small nonelectrolytes. The turgor pressure can be directly measured by means of special pressure probes that can be introduced into the cell (Zimmermann and Neil 1989) or even noninvasively by a leaf patch-clamp pressure probe (Zimmermann et al. 2010).

In contrast to plant cells, animal cells (and also plant protoplasts) do not have the ability to resist an internal hydrostatic pressure. Within certain limits, osmotic changes in the environment can be compensated by alterations in cellular volume, thus maintaining a constant membrane area. Cell swelling, for example, can cause the forming of a sphere or smoothing of the cell surface. As described in Sect. 2.3.4, a mechanical expansion of the cell membrane, without introduction of additional molecules, however, is nearly impossible. An osmotic difference of only 1 mosmol can generate a maximum internal pressure of 2.27 kPa. Measurements made on erythrocyte membranes have shown that their tension can, at the most, only withstand an internal pressure of 0.1 kPa. For this reason, these cells have complicated mechanisms for osmoregulation, including a number of coupled receptors and transport systems.

Consequently, it is not possible to explain the volume regulation of a cell without considering the complete ionic metabolism. We will come back to this problem in connection with the Donnan equilibrium (Sect. 3.2.5).

Further Reading

Osmotically inactive water: Fullerton et al. 2006; Fullerton and Cameron 2007; Garlid 2000; Osmoregulation of animal cells: Okada 2004; Osmotic pressure in plants: Martin et al. 2001; Zimmermann et al. 2004.

3.2.3 Electrochemical Equilibrium: The Nernst Equation

We will now consider a system consisting of two phases, each of which contains a solution with a salt AB, but with different concentrations in phases I and II (Fig. 3.15). Let the salt be completely dissociated into its ions A and B. The membrane is semipermeable in that it allows ion A, but not ion B, to pass through.

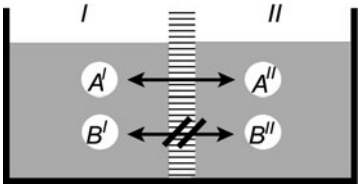

Fig. 3.15 Diagram illustrating the derivation of the Nernst equation. Phases I and II are separated from each other by a membrane which is permeable only for the ion A, but not B, of the salt AB

Before analyzing conditions of thermodynamic equilibrium of this system, we will consider the situation qualitatively. Let us suppose that there is an osmotic equilibration between both phases, compensated by electroneutral components. Thus, alteration of concentrations, induced by volume changes, can be neglected. In this system the electrostatic equilibrium will be disturbed because only ion A, driven by its concentration gradient, but not its counterpart, ion B can penetrate the membrane. This leads to an increase in the electrical potential difference across the membrane. Eventually, a strong electric field hinders a further diffusion of ion A. Ion A, consequently, will be subject to two opposing forces: on the one hand, the driving force, induced by the concentration gradient, i.e., the gradient of its chemical potential and, on the other hand, an opposing electrostatic force which only arises as a result of its own diffusion. The following equilibrium will be established: a few ions cross the membrane, inducing an electric field which stops further diffusion.

The basic condition to calculate this equilibrium is the equality of the electrochemical potentials of ion A between phases I and II:

$$\tilde{\mu}_A^I \stackrel{!}{=} \tilde{\mu}_A^{II}$$

Substituting the expressions for the electrochemical potentials according to Eq. 3.41, one obtains

$$\mu_A^{0I} + \mathrm{R}T \ln a_A^I + z_A \mathrm{F} \psi^I = \mu_A^{0II} + \mathrm{R}T \ln a_A^{II} + z_A \mathrm{F} \psi^{II} \tag{3.110}$$

(We will assume isothermal conditions, i.e., $T^I = T^{\mathrm{II}} = T$.)

In contrast to the derivation of the equation for osmotic pressure, in the present case, the standard potentials of these components of phases I and II are equal, because there is no pressure difference ($\mu_A^{0I} = \mu_A^{0II}$).

Taking this into account, and re-arranging Eq. 3.110, gives

$$z_A \mathrm{F}(\psi^I - \psi^{II}) = \mathrm{R}T(\ln a_A^{II} - \ln a_A^I) \tag{3.111}$$

and therefore

$$\Delta\psi \equiv (\psi^I - \psi^{II}) = \frac{\mathrm{R}T}{z_A \mathrm{F}} \ln \frac{a_A^{II}}{a_A^I} \tag{3.112}$$

This is the *Nernst-Equation*. It gives the electrical potential difference ($\Delta\psi$) across the membrane as a function of the chemical activities of the permeating ion in both phases at thermodynamic equilibrium.

Equation 3.112 can be re-arranged to show the difference of ion activities (a^I and a^{II}) which builds up in two phases, with a given potential difference ($\Delta\psi$) in between:

$$a_A^I = a_A^{II} \mathrm{e}^{-\frac{z_A \mathrm{F} \Delta\psi}{\mathrm{R}T}} \tag{3.113}$$

Such a relation has already been derived and employed using the Boltzmann equation (Sect. 2.1.4), and applied to calculate local ion concentrations near charged particles (Eq. 2.50), or in electric double layers (Eq. 2.77). In these cases, however, the concentrations (c_i) were used instead of the chemical activities (a_i). This is allowed only for ideal, i.e., diluted, solutions, or when the activity coefficients (f_i) are equal in both phases.

The Nernst equation, therefore, permits the calculation on the one hand, of the distribution of ions as a function of the electrical potential (Eq. 3.113) and, on the other hand, the electrical potential, which is induced by an unequal distribution of ions (Eq. 3.112). For both cases, however, thermodynamic equilibrium is required!

The separation of the two phases of the system by a semipermeable membrane, as discussed here, reflects just one special example. In the case of ion distribution in electric double layers, or ionic clouds, as described in Sects. 2.3.5 and 2.3.6, the electric potential gradient is predicted by the fixed charges, and the distributions of the ions are not limited by a membrane. Here, both the anions and the cations in the case of equilibrium are distributed according to this equation.

All equations, derived in the present section, are applicable only for thermodynamic equilibria. This means that the Nernst equation cannot be used to calculate the membrane potential of a living cell. Actually, the membrane potential of a living cell is either a diffusion potential (liquid junction potential), or it is generated by electrogenic ion pumps (see Sects. 3.4.2 and 3.4.3). Conversely, despite the nonequilibrium distribution of the ions in the cell in general, it is quite possible that some types of ions may be distributed passively and then they are actually in equilibrium.

An example of this is the chloride concentration of most animal cells. Because its membrane permeability is rather fast, and because no chloride pumps in the membrane exist, its distribution is mostly passive, and predicted by the existing transmembrane potential. The Nernst equation, therefore, allows one to calculate the internal chloride concentration, if the external chloride concentration and the transmembrane potential are known. Conversely, knowing, for example, the distribution of chloride inside and outside the cells, one can calculate the transmembrane potential of the cell. In this context, however, it is important to underline that in this case the chloride distribution is just an *indicator* for the membrane potential, but not the *reason* for it!

This consideration allows the establishment of a method to measure the transmembrane potential of cells without microelectrodes. The chloride distribution can be easily determined using the radioisotope ^{36}Cl. Sometimes one can use other small charged organic molecules, which penetrate the membrane quickly, and are labeled by ^{3}H or ^{14}C. Knowing the distribution of these ions, the membrane potential ($\Delta\psi$) can be calculated according to Eq. 3.112.

Furthermore, the Nernst equation allows the calculation of electrode potentials. If a metal is dipped into an electrolyte solution, then cations are detached from the

metal lattice and an electrochemical equilibrium is obtained. In general, this is the same situation as discussed above. The resulting electrical potential difference between the metal and the solution is called *electrode potential*. If the cation from the metal of the electrode forms a poorly soluble salt with an anion of the solution, and if the concentration of this anion is much greater than that of the cation, then the electrode potential, linked through the soluble product of the salt, is directly determined by the chemical activity of this anion. In this way, for example, a silver electrode covered by AgCl can be used to measure the Cl^- activity of a solution (see Fig. 3.16). The voltage of this electrode with respect to a reference electrode, according to the Nernst equation (Eq. 3.112), is proportional to the logarithm of the Cl^- activity in the solution. It is important to emphasize that these kinds of electrochemical methods allow the measurement of ion activities (a_i), in contrast to most chemical methods, indicating its concentrations (c_i).

The electrochemical measurement of ion activity and of many other chemical substances has become an important and universal technique. Using special semi-permeable membranes, or water impermeable layers containing special ionophores, electrodes can be made which are highly selective to measure the activity of special chemical components. In this case, these membranes or microphases separate a reference solution of known composition from the test solution. The potential difference between the reference solution and the test solution is measured as the voltage between two reference electrodes dipped in both phases.

A typical example for this kind of measurement is the usual pH electrode (see Fig. 3.16). In this case, a thin membrane of special glass or another material allows the protons to equilibrate between both phases. If, as a reference phase, the pH electrode is filled by a buffer holding the pH constant, then the voltage between the pH electrode and the reference electrode indicates the pH of the test solution. Usually, combinations between pH and reference electrodes are used, consisting of one single glass body.

For electrochemical methods of analysis a pair of electrodes is always necessary. The electromotive force will be measured between a selective measuring electrode and a so-called reference electrode. A *reference electrode* is an electrode that exhibits a constant electrode potential that is independent of the composition of the solution into which it is immersed. If a pair of identical reference electrodes is used, then the electrode potentials will mutually oppose each other and therefore the direct potential difference, i.e., the actual electromotive force between the two phases can be measured. If two different reference electrodes are used, a constant $\Delta\psi$ is superimposed, which can be subtracted from the measured value.

To make a reference electrode, a metal electrode is covered with a slightly soluble salt of this metal. This is immersed into a reference solution of constant composition, which is connected to the test solution by means of a pathway, where little, if any, convection can take place (liquid junction). This liquid junction is formed in different ways. It can be a tube, filled by an agar gel. However, in industrially produced electrodes, it is mostly just an opening in the glass wall, covered by a small filter made from sintered glass. Sometimes a ground glass stopper is used with a liquid film in between.

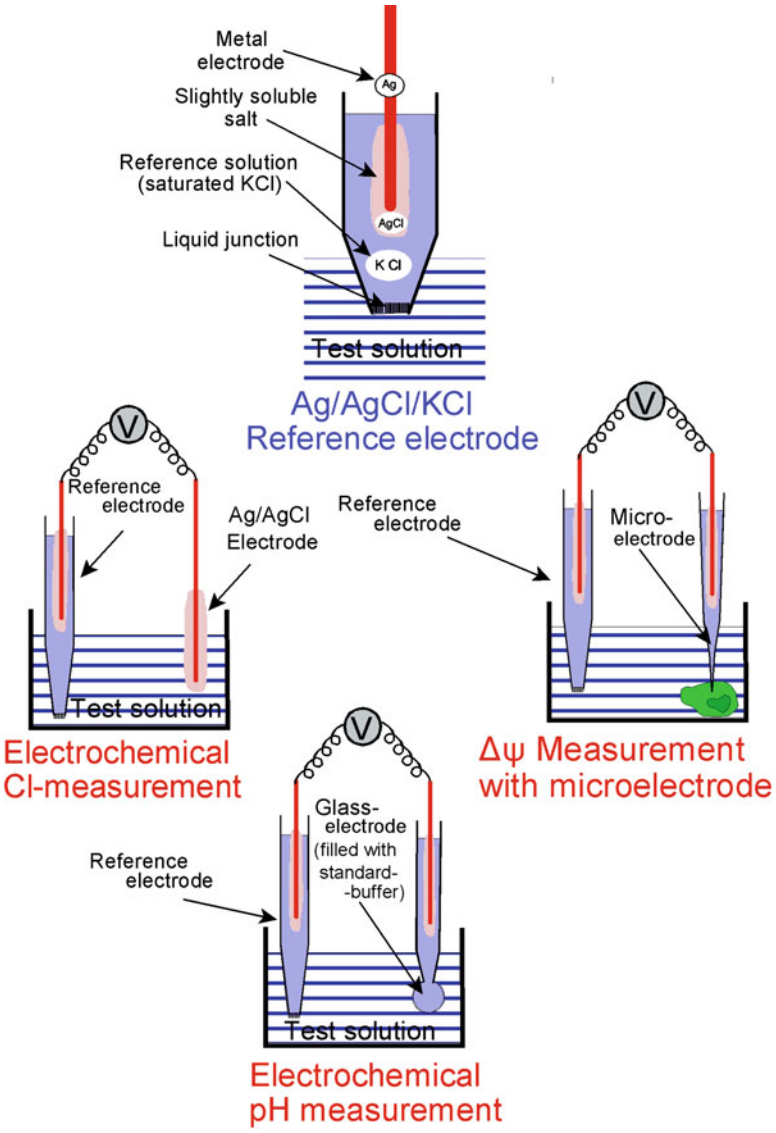

Fig. 3.16 The construction of an Ag/AgCl/KCl-reference electrode, as well as its application for electrochemical measurements of Cl^- activity and pH, and as a reference electrode in electrophysiology

The most common types of reference electrodes are Ag/AgCl/KCl electrodes (see Fig. 3.16). In this case, a saturated solution containing 4.6 M KCl as a reference solution is used. The reason for choosing KCl is that in any case, a concentration gradient exists through the connection between the reference solution and the test solution. This could become a source of an electrical potential difference (diffusion

potential, see Sect. 3.3.3). This would contradict the conditions for reference electrodes, i.e., the constant potential, independent of the test solution. This diffusion potential, however, depends not only on the concentration gradient, but also on the mobility of both ions (see Eq. 3.190). The mobility of the K^+ and Cl^- ions in aqueous solutions, however, is almost equal. This means that even for strong concentration gradients between reference solution and test solution, no diffusion potential can occur. This would not be the case, if for example NaCl was used as the reference solution.

Unfortunately, this useful property of K^+ and Cl^- ions is valid only in pure solutions. If a reference electrode is immersed for example in suspensions of charged particles, a so-called *suspension effect* occurs. Under the influence of the electric double layers of these particles, the mobility of the ions may change. Therefore, small diffusion potentials can occur. This effect, for example is important in the case of pH measurement in blood. Moreover, it can be the source of errors in measurements with microelectrodes in cells.

Microelectrodes, as applied in electrophysiology, are usually Ag/AgCl/KCl electrodes. They are constructed from glass tubes that are drawn out at one end to a very fine capillary (diameter <1 μm). In this case, no further diffusion limitation in the tip is necessary. Because of the electric charge of the glass and the extremely thin glass wall at the tip, so-called *tip potentials* of microelectrodes occur which in worst cases can be as large as several millivolts. Microelectrodes can also be sealed with ion-selective materials so that intracellular ionic activity can be measured directly.

Further Reading

Electrodes in general: Fry and Langley 2005; Varma and Selman 1991; microelectrodes: Amman 1986.

3.2.4 The Donnan Equilibrium: Basic Properties

The Donnan state represents an equilibrium between two phases, containing not only small anions (A) and cations (C), both of which are freely exchangeable between the phases (i.e., can penetrate the membrane), but also charged molecules or particles (M) that are fixed in one phase (i.e., cannot penetrate the membrane; see Fig. 3.17). This type of equilibrium was investigated by F. G. Donnan in 1911.

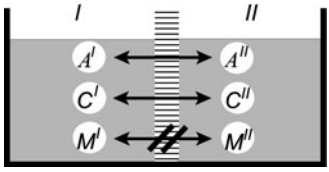

Fig. 3.17 The derivation of the Donnan equilibrium. Phases I and II are separated from each other by a membrane, that is permeable for the anions (A) and cations (C), but not for the charged molecules M

These considerations are of particular importance to understand the properties of various colloidal as well as biological systems where the phases are separated by membranes with particular conditions or permeability.

To analyze this situation, let us denote the exchangeable ions by the index i, and the fixed charge components by m; the concentration, or the activity of the components of one phase, resp. inside the cell with c^I or a^I, and in the second phase, or external solution with c^{II} or a^{II}, respectively. The parameter z_m denotes the number and the sign of charges of the nonexchangeable molecules.

The Donnan equilibrium is defined by the following three conditions:

– All permeable ions (i), being in equilibrium, are distributed according to the Nernst equation (Eq. 3.113):

$$a_i^I = a_i^{II} \mathrm{e}^{-\frac{z_i \mathrm{F} \Delta\psi}{\mathrm{R}T}} \tag{3.114}$$

– In both phases the sum of charges must be zero (electroneutrality condition):

$$\sum z_i c_i + \sum z_m c_m = 0 \tag{3.115}$$

– Water between both phases is distributed according to its thermodynamic equilibrium:

$$\Delta\mu_W = 0 \tag{3.116}$$

This last condition is easily fulfilled in isobaric systems ($\Delta p = 0$) and in the case of free water movement. This applies for animal cells, which to some extent are able to swell or shrink. The volume of plant cells is limited by the rigid cell wall. In this case a so-called *Donnan-osmotic pressure* may occur.

These basic equations can be combined and solved for particular parameters. Equation 3.114 allows us to calculate the relation of the activities of the exchangeable ions inside and outside the cell as a function of the phase potential difference $\Delta\psi$. If the activity of a univalent anion is denoted by a_A ($z_A = -1$), and that of the corresponding cation by a_C ($z_C = +1$), it follows:

$$\frac{a_{\mathrm{A}}^I}{a_{\mathrm{A}}^{II}} = \frac{a_C^{II}}{a_C^I} = \mathrm{e}^{\frac{\mathrm{F}\Delta\psi}{\mathrm{R}T}} \equiv r \tag{3.117}$$

The so-defined parameter r is known as the *Donnan ratio*. According to Eq. 3.117, it is related to the Donnan potential in the following way:

$$\Delta\psi = \frac{\mathrm{R}T}{\mathrm{F}} \ln r \tag{3.118}$$

The Donnan potential ($\Delta\psi$) is substantially determined by the amount of nonexchangeable charged components in the phases, as reflected in condition (3.115).

Let us now consider the situation with just a single kind of cation (C, with $z_C = +1$), and a single kind of anion (A, with $z_A = -1$), and only one kind of charged component (M, with z_M) inside the cell. In this case, the equation of electroneutrality of both phases, according to Eq. 3.115 can be written easily:

$$\begin{cases} c^I_{\mathrm{C}} - c^I_{\mathrm{A}} + z^I_{\mathrm{M}} c^I_{\mathrm{M}} = 0 \\ c^{II}_{\mathrm{C}} - c^{II}_{\mathrm{A}} = 0 \end{cases} \tag{3.119}$$

This system of equations can be solved by reorganizing it and dividing one by the other:

$$\frac{c^I_{\mathrm{A}}}{c^{II}_{\mathrm{A}}} = \frac{c^I_{\mathrm{C}} + z^I_{\mathrm{M}} c^I_{\mathrm{M}}}{c^{II}_{\mathrm{C}}} \tag{3.120}$$

If the activity coefficients of the ions in both phases are equal (i.e., $a_i = c_i$), one can substitute the Donnan ratio (Eq. 3.117) in this equation:

$$r = \frac{1}{r} + \frac{z^I_{\mathrm{M}} c^I_{\mathrm{M}}}{c^{II}_{\mathrm{C}}} \tag{3.121}$$

Re-arranging this, a simple quadratic equation for r is obtained:

$$r^2 - \frac{z^I_{\mathrm{M}} c^I_{\mathrm{M}}}{c^{II}_{\mathrm{C}}} r - 1 = 0 \tag{3.122}$$

which can be solved in the usual way:

$$r = \frac{z^I_{\mathrm{M}} c^I_{\mathrm{M}}}{2c^{II}_{\mathrm{C}}} \pm \sqrt{\left(\frac{z^I_{\mathrm{M}} c^I_{\mathrm{M}}}{2c^{II}_{\mathrm{C}}}\right)^2 + 1} \tag{3.123}$$

The dependence of the Donnan ratio (r) for this simplified case on the expression $\frac{z^I_{\mathrm{M}} c^I_{\mathrm{M}}}{2c^{II}_{\mathrm{C}}}$ is shown in Fig. 3.18. Considering the corresponding definition (3.117), it is clear that the negative values of the Donnan ratio (r) (blue curve) have no real meaning. Therefore, only the positive sign of the root in Eq. 3.123 is of interest. For the case $z^I_{\mathrm{M}}\, c^I_{\mathrm{M}} = 0$ it becomes $r = 1$ and Eq. 3.118 gives $\Delta\psi = 0$. If the nonexchangeable components are negatively charged ($z^I_{\mathrm{M}} < 0$), then $r < 1$. This gives a negative Donnan potential ($\Delta\psi < 0$). Conversely, the value $z^I_{\mathrm{M}} > 0$ means $r > 1$ and therefore $\Delta\psi > 0$.

A reduction of the ionic strength in the external solution, i.e., a reduction of c^{II}_{K}, leads to an increase in the absolute value of the ratio $\frac{z^I_{\mathrm{M}} c^I_{\mathrm{M}}}{2c^{II}_{\mathrm{C}}}$. If $z^I_{\mathrm{M}} < 0$, this means a

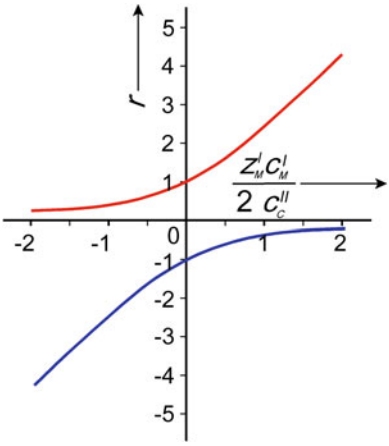

Fig. 3.18 Plot of Eq. 3.123. *Red curve* – positive, *blue curve* – negative values of the root expression

shift of $\Delta\psi$ in a negative direction, and vice versa, if $z^I_{\mathrm{M}} > 0$, a reduction of the ionic strength causes an increase in the Donnan potential. The absolute value of the Donnan potential, therefore, increases when the charge concentration ($z^I_{\mathrm{M}}\, c^I_{\mathrm{M}}$) is increased, as well as when the ionic strength in the external solution (c^{II}_{C}) is reduced.

It should be noted, however, that the Donnan potential can only be measured using reference electrodes with salt bridges (see Fig. 3.16), but not with simple Ag/AgCl electrodes. In fact, if the whole electrochemical system, consisting of reversible electrodes and both electrolyte phases, is in thermodynamic equilibrium, no electromotoric force (emf) can occur. Using electrodes with salt bridges, however, an emf emerges as the difference in the liquid junction potentials at the tops of the reference electrodes that are not in equilibrium. This is called the "indirect method" for determining the Donnan potential. The same is possible by measuring the pH difference between both systems.

Donnan equilibrium not only occurs in phases which are bounded by a membrane but is also of particular importance in various colloidal systems and matrices consisting of charged molecules. It determines the swelling and shrinking of these phases as a result of Donnan-osmotic processes.

Further Reading

Overbeek 1956; Dukhin et al. 2004.

3.2.5 The Donnan Equilibrium in Biological Systems

Although the living system in general and particularly living cells are not in thermodynamic equilibrium, we already pointed out that in fact a number of subsystems nevertheless fulfill equilibrium conditions. Therefore, there exist a number of Donnan systems which are worthy of consideration.

One example concerns the distribution of ions near the fixed charges of the membrane surface coat. This means the extension of the theory of electric double layers (Sect. 2.3.5, Fig. 2.43), to real conditions in the layer of glycoprotein and glycolipid molecules at the surface of most cells (Sect. 2.3.6, Fig. 2.48). Furthermore, this is quite important in order to calculate the conditions of intercellular clefts, and accordingly the Donnan-osmotic behavior of the extracellular space in tissue (Fig. 3.34). In the same way, despite the fact that in most animal cells Na^+ and K^+ ions are actively pumped, others, like Cl^-, as well as water, could be freely exchangeable. Therefore, a Donnan equilibrium occurs as a result of these permeable ions. This also concerns the equilibration of water, resp. volume regulation. In this case, however, not only the intracellular proteins are to be considered as carriers of nonexchangeable charges, but additionally the charges of Na^+ and K^+ are quasi "fixed." Even if the transmembrane potential is determined by diffusion processes (see Sect. 3.3.3), or electrogenic pumps (Sect. 3.4.1), the relation between $\Delta\psi$, the distribution of freely exchangeable ions like Cl^- or pH, and subsequently the volume (V) can be calculated in a similar way.

Furthermore, Donnan equilibrium may occur if the cell membrane is occasionally opened to ions by an ionophore, by a toxin, or by any other influences. A living cell will also shift slowly towards a Donnan equilibrium, if, as a result of low temperature or a chemical blocker, the ATP-driven ion pumps are inhibited. A typical example of this is the distribution of ions in erythrocytes of blood preserves, stored over a longer period of time. These situations occur not only under experimental conditions, but sometimes also in vivo.

Similar Donnan potentials are established in biocolloids like the eye lense or articular cartilage. The fixed charges of the proteoglycans in this case effect the mechanical behavior of articular cartilage by Donnan-osmotic pressure.

In most of these cases the Donnan equilibrium can be considered as a sort of *quasi-equilibrium state* (see Sect. 3.1.4). Figure 3.8 demonstrates the time hierarchy of characteristic rate constants of various biological processes. In human erythrocytes, and in most animal cells, as a rule, the distribution of Cl^- ions between internal and external media is passive. The same is true for the equilibration of pH. These processes of equilibration are very fast, and their stationary state can be calculated according to the Donnan equilibrium. This situation is called a "*quasi*" equilibrium state, because over a longer time interval the concentrations of potassium and sodium may shift. The equilibration of Cl^- concentration, and the pH, therefore follows the slow shift in Na^+ and K^+ concentration, and can be considered as in equilibrium only for a limited time of observation (see Sect. 3.1.4, Fig. 3.8).

Equation 3.120 in the previous chapter allows us to understand the behavior of a Donnan system in general, but in fact, it reflects an extremely simplified situation. It has not been taken into account that according to the condition of iso-osmolarity (Eq. 3.116), changes in cell volume could appear. Volume changes, however, would lead to changes in intracellular concentrations, and consequently in activity and osmotic coefficients of all participants.

Furthermore, according to the Donnan condition, a pH equilibrium between the two phases in this system occurs. A pH change, however, influence the charges of organic molecules. In its simplest form, this dependence can be expressed as:

$$z_M = -z_{M0}(pH - pH_{iso}) \tag{3.124}$$

Near the isoelectric point ($pH = pH_{iso}$) the total molecule is uncharged ($z_M = 0$). Below this point ($pH < pH_{iso}$), z_M becomes positive, and above it ($pH > pH_{iso}$), z_M becomes negative (see Sect. 2.2.8, Figs. 2.30 and 2.31).

Considering all of these relations, one obtains a system of nonlinear equations that can be solved by iteration. Some basic stationary properties of such a feedback system will be demonstrated for the case of Donnan equilibrium in human erythrocytes. As already mentioned, this situation can occur under conditions of blood storage or as a result of other treatment.

The most important nonpenetrating charge component in erythrocytes is hemoglobin. In vivo, it has a concentration of 7 mM. Its isoelectric point is $pH_{iso} = 6.8$ (at 25°C) and $z_{M0} = 10.5$ eq/mol. The Donnan potential ($\Delta\psi$) of erythrocytes in solutions of various NaCl concentrations is shown in Fig. 3.19. The constant osmotic pressure of 300 mosmol of the external solution is maintained by various concentrations of sucrose. The intracellular pH of the erythrocytes depends on the pH value of the external solution which is indicated on the abscissa, as well as on the Donnan potential itself. At the isoelectric point of hemoglobin, the Donnan potential becomes zero ($\Delta\psi = 0$). The highest potentials, both negative and positive, are obtained at the greatest distances from the isoelectric point, and in solutions with the lowest NaCl concentrations.

The volume changes in these cells are shown in Fig. 3.20. The volume, expressed here as the percentage of the in vivo volume of erythrocytes, is indicated as a function of the NaCl concentration of the external solution at constant $pH = 7.4$. It is important to note that independently of the external NaCl concentration, the osmotic pressure of the external medium is always adjusted to the isotonic osmotic pressure:

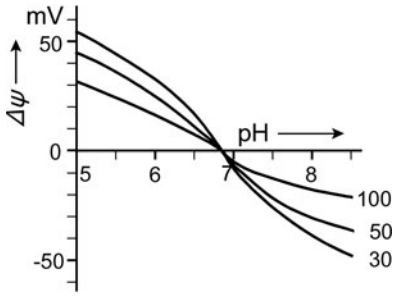

Fig. 3.19 The Donnan potential ($\Delta\psi$) of human erythrocytes in isotonic NaCl-sucrose solutions dependent on external pH. The curves represent situations for NaCl solutions of 30, 50, and 100 mM ($\pi = 300$ mosmol, $T = 25$°C) (After Glaser et al. 1980)

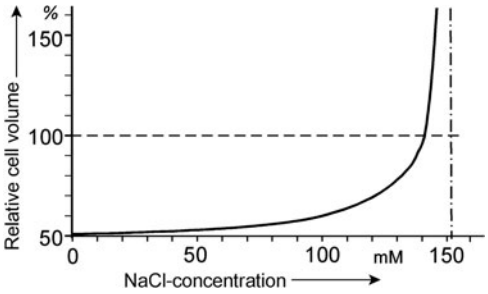

Fig. 3.20 Donnan-osmotic alterations of the relative volume (V) of human erythrocytes as a function of external NaCl concentration in isotonic NaCl-sucrose solutions (pH = 7.4, $T = 25°C$). The osmotic pressure of the solutions with different NaCl-concentrations is always balanced by sucrose, adjusting to $\pi = 300$ mosmol. At $c = 152$ mM, the solution contains only NaCl, without sucrose. The volume is given as the percentage of the in vivo volume of erythrocytes (after Glaser et al. 1980, redrawn)

$\pi = 300$ mosmol. In spite of this, the cells shrink in solutions of low ionic strength. An isotonic condition of the incubation medium, therefore, is no guarantee for the maintenance of a normal cell volume! A normal volume is achieved only in a solution containing about 20 mosmol sucrose and 140 mM NaCl. In this case, the sucrose just compensates the osmotic pressure of the hemoglobin, which has a comparatively high osmotic coefficient (see Fig. 3.13). Furthermore, it is interesting that the volume curve rises steeply at high ionic strengths. Mathematically, no results can be obtained from the equations assuming conditions of pure (isotonic!) 152 mM NaCl solutions. In this case, the volume would become infinitely large. The experiments indicate that erythrocytes in solutions of pure electrolytes undergo hemolysis, if the membrane becomes permeable for these ions. This is known as *Donnan-osmotic hemolysis*. As indicated in Fig. 3.20, this can occur even in isotonic solutions.

To determine experimentally whether a cell is in a state of Donnan equilibrium, the relation between the internal and external ionic activities must be determined. For a Donnan state to be present, the Donnan ratios (r) from Eq. 3.117 must correspond. For most cells in vivo, the ratio obtained using the sum of the sodium and potassium ions $[(a^I_K + a^I_{Na})/(a^{II}_K + a^{II}_{Na})]$ may be close to that for a Donnan equilibrium, but this would not be the case for either Na^+ or K^+ alone. Active transport changes their relative concentrations in opposing directions.

Similar calculations to those demonstrated here for the case of erythrocytes can be applied to other cells, taking into account the membrane potential. In these cases, however, the situation will be complicated by the charges and structure of organelles and the endoplasmatic reticulum net.

Further Reading

Glaser and Donath 1984; Sun et al. 2004; Fraser and Huang 2007.

3.3 Phenomenological Analysis of Fluxes

As mentioned repeatedly, biological functions result in particular molecular processes and appear finally in the form of visible and measurable phenomena. This circumstance becomes apparent especially in various membrane functions as will be described in the next chapters. In a strict sense, the transport of ions and molecules through biological membranes must be considered as a highly specific process of molecular interaction of these species with the structure of a particular transport protein (see Sect. 3.4.5). Nevertheless, the classical approaches based on phenomenological thermodynamic considerations have been used extensively and with considerable success to investigate various processes of molecular and ion transport, and their consequences in volume regulation and various electric phenomena.

A flux, as defined phenomenologically in Sect. 3.1.3, is the amount of a substance, which passes in a perpendicular direction through a definite area of a surface in a unit of time. Its basic measure therefore, is: kg s^{-1} m^{-2}, or better: mol s^{-1} m^{-2}. In the following discussion, the molar flux of a substance will be denoted by the symbol **J**.

When considering fluxes through cell membranes difficulties may arise in some cases if the exact surface area of a cell is not known. In these cases modified units are used, such as: "mol s^{-1} per cell," "mol s^{-1}per liter of cells," "mol s^{-1} per liter of cell-water," etc. As the unit of time, instead of seconds, minutes, hours, or even days are frequently used. Such units are quite convenient for some physiological investigations. However, it is not possible to substitute such units directly into the thermodynamic equations.

3.3.1 The Flux of Uncharged Substances

Let us first consider the diffusion of an uncharged substance (i), not considering any coupling with other fluxes. Using Eqs. 3.45, 3.49, and 3.33, gives

$$\mathrm{J}_i = L_i \mathrm{X}_i = -L_i \ \mathrm{grad}\ \mu_i = -L_i \ \mathrm{grad}\ (\mu_i^0 + \mathrm{R}T \ln a_i) \tag{3.125}$$

In order to calculate the gradient of the chemical potential the dependence of the parameters on their position in space must be considered. Assuming constant temperature (grad $T = 0$), and constant pressure (grad $p = 0$), the gradient of the standard potential (grad μ_i^0) also becomes zero. Equation 3.125 therefore can be written as:

$$\mathrm{J}_i = -L_i \, \mathrm{R}T \ \mathrm{grad}\ \ln a_i = -\frac{L_i \mathrm{R}T}{a_i} \mathrm{grad}\ a_i \tag{3.126}$$

(To explain this rearrangement: the differential operator "grad" can be handled like a deviation d/dx. Then the rules of sequential differentiation are applied).

As has already been discussed in Sect. 3.1.3 a flux can also be expressed by the parameters concentration (c_i) and velocity ($\mathbf{v}_i$) (Eq. 3.47). From this we came to an equation which includes the mobility (ω_i) (Eq. 3.53). If we consider, furthermore, that $\mathbf{X}_i$ is not a simple mechanical force, but results from the gradient of a chemical potential determined by molar concentrations, the Avogadro number (N) must be included.

$$\mathrm{J}_i = \frac{c_i \omega_i}{\mathrm{N}} \mathrm{X}_i \tag{3.127}$$

Comparing Eq. 3.125 and 3.127 results in:

$$L_i = \frac{c_i \omega_i}{\mathrm{N}} \tag{3.128}$$

Suppose $c_i \approx a_i$, then the combination of Eq. 3.128 with 3.126 gives:

$$\mathrm{J}_i = -\frac{\omega_i}{\mathrm{N}} \mathrm{R}T \text{ grad } c_i = -\omega_i \mathrm{k}T \text{ grad } c_i \tag{3.129}$$

Introducing the diffusion coefficient one gets *Fick's first law of diffusion*:

$$\mathbf{J}_i = -D_i \text{ grad } c_i \tag{3.130}$$

If there is only a one-dimensional concentration gradient in the x-direction this equation simplifies to:

$$\mathbf{J}_{ix} = -D_i \frac{\mathrm{d}c_i}{\mathrm{d}x} \tag{3.131}$$

The mobility (ω_i) of an almost spherical molecule can be interpreted mechanically by referring to Stoke's law (Eq. 3.54). By definition it has the unit: m s^{-1} N^{-1}. In Sect. 2.1.6 we introduced the diffusion coefficient (D_i) in the context of the translational movement of molecules. This led to relations between the diffusion coefficient, mobility (ω_i), and the molar mass (M_i) (see Eqs. 2.38–2.40).

We introduced the diffusion coefficient as $D_i = \omega_i kT$. Its unit therefore is $\mathrm{m}^2\ \mathrm{s}^{-1}$. Using the measure mol m^{-3} m^{-1} = mol m^{-4} for the concentration gradient in Eq. 3.130 or Eq. 3.131, one obtains the unit mol s^{-1} m^{-2} for the flux, according to its definition.

In addition to Fick's first law of diffusion, which gives the flux as a function of the concentration gradient, *Fick's second law* allows one to determine the establishment of a concentration gradient of a given substance (c_i) as a function of time (t).

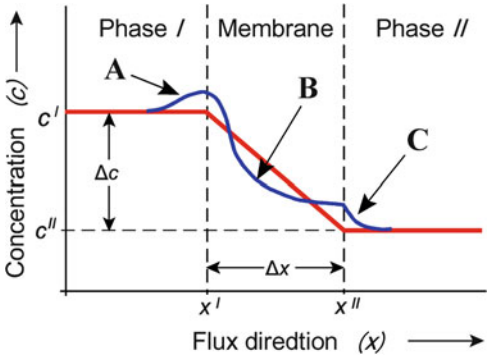

Fig. 3.21 Possible functions $c(x)$ in a membrane system. *Red* – ideal case, *blue* – disturbed case, caused by the following reasons: (**a**) Adsorption of substance at the membrane surface; (**b**) differences in the mobility of the substance inside the membrane; (**c**) diffusion layer caused by imperfect stirring at the surface

It is a partial differential equation of the second order which for diffusion in one direction is:

$$\left(\frac{\partial c_i}{\partial t}\right)_x = D\left(\frac{\partial^2 c_i}{\partial x^2}\right)_t \tag{3.132}$$

This equation is used to calculate a concentration gradient which occurs when diffusion takes place in a homogeneous phase.

In contrast to this case where a continuous gradient of concentration occurs, in membrane systems various discontinuities in concentration are to be expected. This means that the function $c_i(x)$ could become rather complicated. Schematically, this is illustrated in Fig. 3.21. The simplest case is shown by the red line where phase I contains a solution with concentration c^I and correspondingly, c^{II} is the concentration in phase II. The concentration in the membranes falls linearly with a slope of $\Delta c/\Delta x$.

The blue line in Fig. 3.21 shows an irregular concentration pattern. In this case effects are considered which can occur at the membrane surface as well as in its interior. The deviation **A** marks an adsorption of the substance at the membrane surface, or a change in concentration of ions in an electric double layer. The effective thickness of this layer is very small, being less than 10 nm. Inside the membrane (Fig. 3.21B), deviations of the linear concentration profile can occur by differences of the mobility of the substance in the x-direction or, in the case of ion transport, even by dielectric inhomogeneities.

In special cases *diffusion layers,* also called *unstirred* or *boundary layers*, may occur near the surface of membranes (Fig. 3.21C). These are near membrane regions without streaming and convections. In this region substances transported through the membrane, or involved in a chemical reaction, can move only by diffusion. An increase or decrease of local concentration can occur, which depends on the relationship between the transport or reaction rate, introducing substance into the region, and the rate of its removal, i.e., the rate of diffusion. In this case a stationary concentration gradient is built up.

In systems with artificial ion exchange membranes, large diffusion layers can be indicated by special interference methods or by microelectrodes. In contrast to

speculation in earlier papers, in vivo such layers are mostly a lot less than 1 μm. These layers may significantly affect biochemical reactions or transport processes of biological membranes. They can become important especially in cases where the cell surface is covered by microvilli or special caverns, or where reactions take place in the intercellular space (Fig. 3.34). Even the occurrence of diffusional layers of protons near membranes is discussed.

Let us now consider the simplest case, represented by the solid line in Fig. 3.21 in more detail. Let the concentration gradient (dc_i/dx) inside the membrane be constant and equal $\Delta c_i/\Delta x$, whereas: $\Delta c_i = c_i^{II} - c_i^{I}$. In this case Eq. 3.131 becomes

$$\mathbf{J}_i = -D_i \frac{\Delta c_i}{\Delta x} \equiv -P_i \Delta c_i \tag{3.133}$$

The parameter $P_i = D_i/\Delta x$ is the *permeability coefficient* measured in m s^{-1}. The same parameter will be used in Sect. 3.3.3 to calculate fluxes of ions. It is important to stress that this is the same parameter with an identical definition.

Let us now consider a system with flux interactions. In this case the flux coupling must be taken into account as discussed in Sect. 3.1.3. To demonstrate these approaches, we will consider only the simplest case of a binary system, represented for example by a flux of an uncharged substance ($\mathbf{J}_s$) and its solvent, the flux of water ($\mathbf{J}_w$). The driving forces of both fluxes are the negative gradients of their chemical potentials. To simplify the derivation we will use simply their differences $\Delta\mu_s$ and $\Delta\mu_w$ as driving forces.

In a first step we must write down the equation for the dissipation function according to the rule of Eq. 3.64:

$$\Phi = \mathbf{J}_w \Delta\mu_w + \mathbf{J}_s \Delta\mu_s \tag{3.134}$$

In this case $\mathbf{\Phi}$ is an integral parameter for the whole membrane thickness Δx.

In the next step we will modify this equation in such a way that instead of the parameters $\Delta\mu_s$ and $\Delta\mu_w$, forces are introduced that are directly measurable.

Let us first consider the difference of chemical potential of the water ($\Delta\mu_w$).

$$\Delta\mu_w = \Delta\mu_{wx}^0 + \mathrm{R}T \ln \frac{x_w^I}{x_w^{II}} \tag{3.135}$$

In Sect. 3.2.1 we discussed the chemical potential of water in detail and considered in particular its dependence on mole fraction (x_w) as well as on pressure (p). Now we will make use of these derivations. Using Eqs. 3.88 and 3.89, as well as the definition of osmotic pressure, we come to:

$$\Delta\mu_w^0 = \bar{V}_w \Delta p \quad \text{and:} \quad \mathrm{R}T \ln \frac{x_w^I}{x_w^{II}} = -\bar{V}_w \Delta\pi \tag{3.136}$$

where $\bar{V}$ is the partial volume of water, Δp – the difference of hydrostatic pressure, and $\Delta\pi$ – the difference of osmotic pressure. Equation 3.135 therefore can be rewritten as

$$\Delta\mu_w = \bar{V}_w(\Delta p - \Delta\pi) \tag{3.137}$$

Let us now consider the other chemical potential difference in Eq. 3.134, namely $\Delta\mu_s$. This parameter depends on pressure difference in the same way as the difference of the chemical potential of water. According to Eqs. 3.135 and 3.136 one can write:

$$\Delta\mu_s = \bar{V}_s\Delta p + \mathrm{R}T\ln\frac{c_s^I}{c_s^{II}} \tag{3.138}$$

The second term of this equation can be expanded as a series using the common rule for parameters $x > 0$:

$$\ln x = 2\left[\left(\frac{x-1}{x+1}\right) + \frac{1}{3}\left(\frac{x-1}{x+1}\right)^3 + \frac{1}{5}\left(\frac{x-1}{x+1}\right)^5 + \ldots\right] \tag{3.139}$$

Now we substitute for x the value c^I/c^{II}. With good approximation we are justified in retaining only the first term of this series:

$$\ln\frac{c^I}{c^{II}} = 2\left(\frac{\frac{c^I}{c^{II}} - 1}{\frac{c^I}{c^{II}} + 1}\right) = \frac{\frac{c^I - c^{II}}{c^I + c^{II}}}{2} = \frac{\Delta c}{\bar{c}} \tag{3.140}$$

This shows that the logarithm of the ratio of the two concentrations, or even activities, can be replaced by the difference in concentration (Δc) divided by the arithmetic mean ($\bar{c}$) of the concentrations of the two phases. This connection is not only true approximately, as seen by this expansion of the series, but it can be proved mathematically that it is exactly equal.

Relation (3.140), applied to Eq. 3.138, together with the Van't Hoff equation for osmotic pressure: $\Delta\pi = \mathrm{R}T\Delta c$ (Eq. 3.99), gives

$$\Delta\mu_s = \bar{V}_s\Delta p + \frac{1}{\bar{c}_s}\Delta\pi \tag{3.141}$$

Now we have really found reasonable expressions for the differences of the chemical potentials, and using them we can rearrange the equation for the dissipation function (Eq. 3.134). Introducing Eqs. 3.138 and 3.141 into Eq. 3.134 we get:

$$\Phi = J_w\bar{V}_w(\Delta p - \Delta\pi) + J_s\left(\bar{V}_s\Delta p + \frac{1}{\bar{c}_s}\Delta\pi\right) \tag{3.142}$$

and after rearrangement:

$$\Phi = \Delta p(J_w \bar{V}_w + J_s \bar{V}_s) + \Delta\pi\left(\frac{J_s}{\bar{c}_s} - J_w \bar{V}_w\right) \tag{3.143}$$

The expressions in parentheses can be regarded as new flux variables. Flux was defined as the amount of a substance in moles that traverses a surface in a unit of time. Multiplying this parameter by the partial molar volume ($\bar{V}$), one obtains a volume flow. Let us define a total volume flux ($\mathbf{J}_V$) as the sum of individual volume fluxes of all components:

$$\mathbf{J}_V = \mathbf{J}_w \bar{V}_w + \mathbf{J}_s \bar{V}_s \tag{3.144}$$

To illustrate the meaning of the second term of Eq. 3.143, let us remember the relation $\mathbf{J}_i = c_i \mathbf{v}_i$ (Eq. 3.47). The term: $\mathbf{J}_S/\bar{c}_s$ in Eq. 3.143 therefore is an expression for the velocity of the substance ($\mathbf{v}_s$). The volume flow of the solvent ($\mathbf{J}_w\ \bar{V}_w$) can also be considered as the velocity of it ($\mathbf{v}_w$). From this it follows:

$$\frac{\mathbf{J}_s}{\bar{c}_s} - \mathbf{J}_w \bar{V}_w = \mathbf{v}_s - \mathbf{v}_w \equiv \mathbf{J}_D \tag{3.145}$$

The parameter $\mathbf{J}_D$, called *exchange flux* is therefore the difference between the velocity of solute relative to the solvent.

As a result of these transformations we finally obtain a dissipation function which contains measurable parameters:

$$\Phi = \mathbf{J}_V \Delta p + \mathbf{J}_D \Delta\pi \tag{3.146}$$

This makes it possible to write down a flux matrix following the pattern of Eq. 3.58:

$$\begin{aligned} \mathbf{J}_V &= L_V\ \Delta p + L_{VD}\Delta\pi \\ \mathbf{J}_D &= L_{DV}\Delta p + L_D \Delta\pi \end{aligned} \tag{3.147}$$

This equation shows that in the case of the transport of a solution with a single uncharged substance, the membrane permeability is determined by four coefficients: L_V, L_{VD}, L_{DV}, and L_D. The driving forces are the osmotic ($\Delta\pi$) and hydrostatic (Δp) pressures.

Some of the coefficients used in these phenomenological equations can easily be illustrated: The parameter L_V shows for example how fast the solution passes through the membrane in response to a hydrostatic pressure difference (Δp). Introducing the conditions $\Delta\pi = 0$ and $\Delta p > 0$ in the upper part of Eq. 3.147 it becomes: $L_V = \mathbf{J}_V/\Delta p$. This means that L_V is a mechanical *filtration coefficient* or a kind of *hydraulic conductivity*.

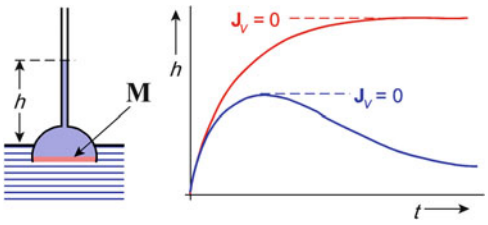

Fig. 3.22 The height of the water column in a Pfeffer osmometer dependent on time in two experiments under different conditions: Pfeffer's cell is closed by a semipermeable membrane (**M**) (*red line*), and: solvent as well as solute can penetrate the membrane (*blue line*). The conditions with stationary pressure in this manometer ($dh/dt = 0$, i.e., $dp/dt = 0$) mean at the same time stationary volume ($dV/dt = \mathbf{J}_V = 0$)

Under the same conditions the substance can also be forced through a filter. In this case the second line in Eq. 3.147 gives $L_{DV} = \mathbf{J}_D/\Delta p$. Because of this, L_{DV} is called the *ultrafiltration coefficient*.

The flux matrix (3.147) makes it possible to describe the time dependence of an osmotic system. Figure 3.22 shows how the pressure in a Pfeffer's cell, measured by the height of the water column in the vertical tube changes with time. Only in the case of a semipermeable membrane (red line), will a thermodynamic equilibrium with a constant pressure difference be achieved. If in addition to the solvent the membrane allows some of the solute to pass through then there will be a decline in the pressure after the initial rise. The osmotic pressure inside the osmometer will continuously fall (blue line). In this case a state without any volume flux ($dV/dt = \mathbf{J}_V = 0$) is achieved only for a short time. For this situation it follows from the first line of the flux matrix (Eq. 3.147):

$$L_V \Delta p + L_{VD} \Delta \pi = 0 \tag{3.148}$$

and furthermore:

$$(\Delta p)_{J_V=0} = -\frac{L_{VD}}{L_V} \Delta \pi \tag{3.149}$$

In Sect. 3.2.2 we introduced Staverman's reflection coefficient (σ). Relating Eqs. 3.108 to 3.149, the following connection between the reflection coefficient and the coupling coefficients results:

$$\sigma = -\frac{L_{VD}}{L_V} \tag{3.150}$$

For a semipermeable membrane with: $\sigma = 1$, it therefore holds that:

$$L_{VD} = -L_V$$

A better illustration of this parameter allows the following consideration: A solution will be forced through a membrane by a hydrostatic pressure ($\Delta p > 0$) without any osmotic difference ($\Delta\pi = 0$). Using these conditions and dividing the second equation of the flux matrix (3.147) by the first one, we obtain:

$$\frac{J_D}{J_V} = \frac{L_{DV}}{L_V} = -\sigma \tag{3.151}$$

According to Eq. 3.145, $\mathbf{J}_D$ can be replaced by the difference ($\mathbf{v}_s - \mathbf{v}_w$). For very dilute solutions it holds: $\mathbf{J}_w\overline{\mathrm{V}}_w \gg \mathbf{J}_s\overline{\mathrm{V}}_s$. In this case Eq. 3.151 can be written as:

$$\sigma = \frac{\mathbf{v}_w - \mathbf{v}_s}{\mathbf{v}_w} \tag{3.152}$$

This equation has already been discussed in Sect. 3.2.2 (Eq. 3.109).

The deviations shown here clearly demonstrate that the consideration of more complex systems, for example solutions with more components, would lead to a huge number of parameters and relations.

This theory of coupled fluxes has been widely used to explain the stationary volume and stationary pressure observed in experiments with cells under nonequilibrium conditions. Such experiments have been carried out mostly on red blood cells and on plant cells. Human red blood cells, placed initially in an isotonic solution, shrink immediately when the osmotic pressure of the external solution is increased by adding a permeable substance i ($\sigma_I < 1$). Then however, if the substance passes through the membrane, the cell usually reverts to its initial volume, as illustrated in Fig. 3.22 by the blue line. Such experiments with fast volume changes are undertaken to determine the reflection coefficient using stop-flow techniques.

Further Reading

Katchalsky and Curran 1965; Stein 1990; Zimmermann and Neil 1989. Papers on unstirred diffusion layers: Barry 1998; Evtodienko et al. 1998; Pohl et al. 1998.

3.3.2 Fluxes of Electrolytes

The diffusion of ions is governed by the same fundamental laws as fluxes of uncharged substances (see Eq. 3.125).

$$\mathbf{J}_i = L_i\mathrm{X}_i = \frac{c_i\omega_i}{\mathrm{N}}\mathrm{X}_i = -\frac{c_i\omega_i}{\mathrm{N}}\,\mathrm{grad}\;\tilde{\mu}_i \tag{3.153}$$

In the case of electrolyte fluxes the driving force is induced by the negative gradient of the electrochemical potential ($\tilde{\mu}_i$). The coupling coefficient (L_i) again can be considered as a product of concentration (c_i) and mobility (ω_i/N).

Let us first consider the expression grad $\tilde{\mu}_i$. For a concentration gradient only in the x-direction, instead of the differential operator *grad*, the differential quotient can be applied:

$$\frac{\mathrm{d}\tilde{\mu}_i}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}x}(\mu_i^0 + \mathrm{R}T \ln a_i + z_i \mathrm{F}\psi) \tag{3.154}$$

In order to simplify the equation, let the solution be close to ideal, i.e., let the activity coefficient be: $f_i \approx 1$, and therefore $a_i \approx c_i$. Furthermore, let the system be under isobaric (grad $p = 0$), and isothermic (grad $T = 0$) conditions. In this case Eq. 3.154 becomes

$$\frac{\mathrm{d}\tilde{\mu}_i}{\mathrm{d}x} = \frac{\mathrm{R}T}{c_i}\frac{\mathrm{d}c_i}{\mathrm{d}x} + z_i \mathrm{F}\frac{\mathrm{d}\psi}{\mathrm{d}x} \tag{3.155}$$

which when combined with Eq. 3.153, results in

$$\mathbf{J}_i = -\frac{c_i \omega_i}{\mathrm{N}}\left(\frac{\mathrm{R}T}{c_i}\frac{\mathrm{d}c_i}{\mathrm{d}x} + z_i \mathrm{F}\frac{\mathrm{d}\psi}{\mathrm{d}x}\right) \tag{3.156}$$

or modified using the diffusion coefficient according to Eqs. 3.129, 3.130

$$\mathbf{J}_i = -D\left(\frac{\mathrm{d}c_i}{\mathrm{d}x} + \frac{z_i \mathrm{F} c_i}{\mathrm{R}T}\frac{\mathrm{d}\psi}{\mathrm{d}x}\right) \tag{3.157}$$

This is the *Nernst–Planck equation*. It contains the differential quotients of the concentration $[c_i(x)]$, and of the electrical potential $[\psi(x)]$. These differential quotients can be integrated only if the corresponding functions are known. This problem has already been discussed with regard to the concentration $[c_i(x)]$ in Sect. 3.3.1 (see Fig. 3.21). In contrast to the case of Fick's equation, here the function $\psi(x)$ must be known in addition (see Fig. 2.48).

The simplest approach is the consideration of linear gradients. This corresponds for example to a membrane with large, water-filled, and noncharged pores where the ions can move freely as in bulk water. Integrating Eq. 3.157 for these conditions, instead of the differential quotients, simple ratios of differences appear, and instead of concentration c_i the mean concentration of both phases $[\bar{c}_i = (c^I + c^{II})/2]$ appears.

$$\mathrm{J}_{ix} = -D\left(\frac{\Delta c_i}{\Delta x} + \frac{z_i \mathrm{F}\bar{c}_i}{\mathrm{R}T}\frac{\Delta\psi}{\Delta x}\right) \tag{3.158}$$

Or, using the permeability coefficient: $P_i = D/\Delta x$ (see Eq. 3.133):

$$\mathrm{J}_{ix} = -P_i\left(\Delta c_i + \frac{z_i \mathrm{F}\bar{c}_i}{\mathrm{R}T}\Delta\psi\right) \tag{3.159}$$

In 1943, D. E. Goldman integrated the Nernst–Planck equation, supposing only the so-called *constant field conditions*, i.e., assuming: $\mathbf{E} = -\mathrm{grad}\ \psi = \mathrm{const.}$

The concentration profile results from the bulk concentrations in both phases, and from its passive distribution in the electric field. Integrating the Nernst–Planck equation (3.157) with these conditions, one gets the following expression:

$$\mathbf{J}_i = -P_i\beta\frac{c_i^I - c_i^{II}e^{\beta}}{1 - e^{\beta}} \qquad \text{with :} \qquad \beta = \frac{z_i F}{RT}\Delta\psi \tag{3.160}$$

The function $\mathbf{J} = f(\Delta\psi)$ is illustrated in Fig. 3.23. It considers the flux of a monovalent cation ($z_i = +1$) penetrating the membrane with a permeability $P_i = 10^{-7}$ ms^{-1}. Let the flux ($\mathbf{J}_i$) be positive if it is directed phaseI $\Rightarrow$ phaseII. A negative $\mathbf{J}_i$ therefore, means that the flux is in the opposite direction. The potential gradient is negative per definition, if ψ decreases from phase I to phase II.

The red line in Fig. 3.23 represents a cation flux which is driven only by the electric field without a concentration difference ($c_i^I = c_i^{II} = 100$ mM). In this case, the flux vanishes ($\mathbf{J}_i = 0$) if there is no potential difference ($\Delta\psi = 0$).

If there is an additional driving force resulting from a concentration gradient, the curve for example is displaced towards one of the two blue lines. In these cases an ion flux exists even if $\Delta\psi = 0$. If $c_i^I = 100$ mM, $c_i^{II} = 10$ mM, and $\Delta\psi = 0$, there is a flux from I to II (i. e. $\mathbf{J}_i > 0$). When the concentrations are reversed, $\mathbf{J}_i$ becomes negative according to the given definition. The blue lines cut the abscissa at +60 mV and −60 mV, respectively. These points mark where the driving force due to the electric field exactly balances the driving force due to the concentration gradient. This corresponds to the equilibrium situation with $\mathbf{J}_i = 0$, described by the Nernst equation (Eq. 3.113).

In fact, the Goldman equation is a good approach even in cases, where the linearity of the function $\psi(x)$ is not exactly realized. This equation is used frequently in various physiological calculations.

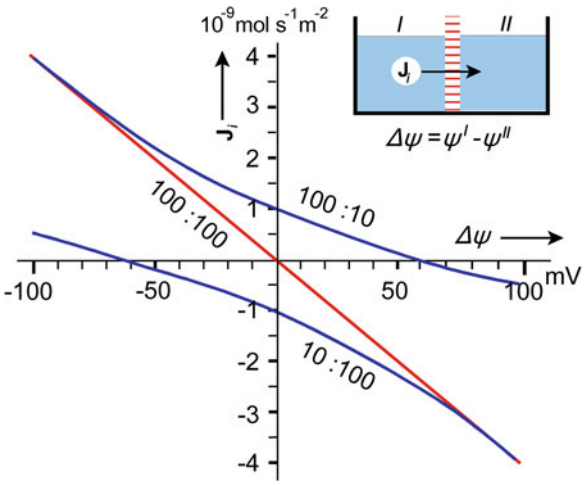

Fig. 3.23 The flux ($\mathbf{J}_i$) of a univalent cation ($z_i = +1$, $P_i = 10^{-7}$ ms^{-1}) driven by a concentration, as well as an electrical potential gradient according to the Goldman equation (3.160). The concentrations in phases I and II are listed on the curves. Description in the text

Furthermore, in recent calculations electrostatic interactions of the ions being transported have been taken into account, as well as the fixed charges of the pores. For this, the Nernst–Planck approach is combined with the Poisson equation, predicting the field divergency ($\nabla^2\psi$) as a function of the three-dimensional charge distribution $\rho(x,y,z)$. In this case the generalized Poisson equation (see Sect. 2.2.4, Eqs. 2.51–2.53) must be written as:

$$\nabla^2\psi = \frac{1}{\varepsilon\varepsilon_0}\left(\mathrm{F}\sum_{i=1}^{n} c_i z_i + \rho\right) \tag{3.161}$$

where the sum term gives the charge density of mobile ions in the electrolyte, and ρ represents the fixed charges on the boundary of the pore. This expression can be combined with Eq. 3.157, written also in the general form:

$$\mathbf{J}_i = -D\left(\nabla \mathrm{C}_i + \frac{z_i \mathrm{F} c_i}{\mathrm{R}T}\nabla\psi\right) \tag{3.162}$$

This approach, known as *Poisson–Nernst–Planck theory* (PNP-Theory), can be solved analytically only in some very specialized cases. Using it to calculate transport in discrete ion channels, numerical computer simulations are required.

In any case, the following limitations of these approaches must be taken into account:

- In fact knowledge of the appropriate dielectric constant (ε) (see Eq. 3.161) for channels has not as yet been established. For calculation of the molecular structures of proteins and lipids, usually the dielectric constant of $\varepsilon \sim 2$ is applied, in electrolyte medium that of water: $\varepsilon \sim 80$.
- Special dielectric effects, such as dehydration of the ions inside the channels, specific short-range interactions of ions with polar amino-acid side chains at these locations, etc., are not taken into account.
- The problem is reduced to a steady-state situation. Ion permeation in channels, however, is basically a time-dependent process. As an ion moves through the channel, the forces acting on it change.

In addition to these new approaches, a "classical" one should not be forgotten: In 1949 Hans Ussing introduced an equation which has been used even in recent physiological publications. It makes it possible to calculate the ratios of ionic fluxes independently of the functions $c(x)$ and $\psi(x)$.

In this context a term must be defined that we will need in further considerations: The flux that is measurable directly through changes of chemical concentrations is called the *net flux* ($\mathbf{J}$). Using radioactive isotopes it is possible to indicate that this net flux in fact results from the difference of two opposite *unidirectional fluxes* $\mathbf{J}_{12}$ and $\mathbf{J}_{21}$:

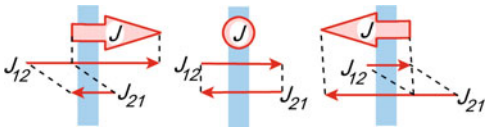

Fig. 3.24 The net flux (**J**), as the sum of the unidirectional fluxes $\mathbf{J}_{12}$ and $\mathbf{J}_{21}$

$$\mathbf{J} = \mathbf{J}_{12} - \mathbf{J}_{21} \tag{3.163}$$

In contrast to the unidirectional fluxes, the direction of which is indicated by the subscripts (for example: $\mathbf{J}_{12}$, $\mathbf{J}_{21}$), the net flux will be defined as positive, as in the example of Fig. 3.24, if it takes place in the direction corresponding to Eq. 3.163.

It must be noted that these are examples of simple diffusion processes based on thermodynamic fluctuations. Later in Sect. 3.4.1 (see Fig. 3.25) we will consider systems of co-transport through biological membranes, which shows the same kinetics but which must be treated thermodynamically with completely different approaches. These two kinds of transport processes therefore must not be mixed.

Considering Eq. 3.163, one can write Fick's equation (Eq. 3.133) in the following way:

$$\mathbf{J} = \mathbf{J}_{12} - \mathbf{J}_{21} = -P\ \Delta c = P(c^I - c^{II}) = P\ c^I - P\ c^{II} \tag{3.164}$$

From this we can conclude with certain justification:

$$\mathbf{J}_{12} = Pc^I; \qquad \mathbf{J}_{21} = Pc^{II} \tag{3.165}$$

This, by the way, resembles the approach of compartment analysis and reactions of the first order as will be used in Sect. 5.1.1.

To calculate unidirectional fluxes of ions one could use the same approaches as in Eq. 3.165, but just introducing a kind of electrochemical concentration ($\tilde{c}$). This parameter, as we will indicate later, has no further relevance. The definition results from the following equation:

$$\tilde{\mu} = \mu^0 + \mathrm{R}T \ln c + z\mathrm{F}\psi \stackrel{!}{=} \mu^0 + \mathrm{R}T \ln \tilde{c} + z\mathrm{F}\psi^0 \tag{3.166}$$

Therefore: $\tilde{c} = c$, if $\psi^0 = \psi$. This means that a kind of zero-potential will be established, whatever it is. From this definition follows:

$$\ln \tilde{c} = \ln c + \frac{z\mathrm{F}}{\mathrm{R}T}(\psi - \psi^0) \tag{3.167}$$

and:

$$\tilde{c} = c\, \mathrm{e}^{\frac{zF}{\mathrm{R}T}(\psi - \psi^0)} \tag{3.168}$$

This equation is of little use since the reference potential ψ^0 is unknown. However, if this relation is substituted in Eq. 3.165, and if the ratio of the unidirectional fluxes are calculated, all unknown parameters cancel:

$$\frac{\mathbf{J}_{12}}{\mathbf{J}_{21}} = \frac{c^I}{c^{II}} \; e^{\frac{zF}{RT}(\psi^I - \psi^{II})} \tag{3.169}$$

(To avoid the accumulation of subscripts we ignored in these derivations the index for the specific substance i).

This formula is known as *Ussing's equation* or *Ussing's flux ration criterion* which relates unidirectional fluxes to concentrations and electrical potential differences. All parameters of this equation can be measured and the validity of the equation therefore, can be checked experimentally. As preconditions for this equation only differences of concentrations and electrical potentials are considered. If the relation between two measured unidirectional fluxes does not agree with the results calculated by these gradients, then additional driving forces are involved. This could be a hint at coupling processes or of any kind of active transport processes.

Further Reading

Goldman 1943; Kontturi et al. 2008; Roux et al. 2004; Syganov and von Klitzing 1999; Ussing 1949; Zhou and Uesaka 2009.

3.3.3 The Diffusion Potential

The diffusion of an electrolyte, in general, can be considered as a separate movement of the dissociated ions along their particular electrochemical gradient, coupled, however, with the electric field resp. the potential difference $\Delta\psi$. Cations and anions, however, may have different mobilities (w_i). The slower diffusion types of ions will lag behind the faster ones. This produces a local potential difference called the *diffusion potential*, retarding the faster ions and speeding up the slower ones. Diffusion potentials can occur in homogenous solutions as well as between two phases separated by a membrane which is permeable for the ions.

An equation for the diffusion potential can be derived, postulating the electroneutrality of the sum of all ion fluxes. In this case the cation flux ($\mathbf{J}_c$) induces a flow of charge ($\mathbf{J}_c z_c$). In the case of electroneutrality, it must be compensated by the charge transport of anions ($\mathbf{J}_A z_A$):

$$\mathbf{J}_C z_C + \mathbf{J}_A z_A = 0 \tag{3.170}$$

Now, we can use the flux equations derived in the previous Sect. 3.3.2. If linear functions $c(x)$ and $\psi(x)$ are proposed, the simplest flux equations (Eq. 3.159) can be used. Inserting them into Eq. 3.170, one obtains:

$$z_C P_C \Delta c_C + \frac{z_C^2 \mathrm{F} \bar{c}_C P_C}{\mathrm{R}T} \Delta\psi + z_A P_A \Delta c_A + \frac{z_A^2 \mathrm{F} \bar{c}_A P_A}{\mathrm{R}T} \Delta\psi = 0 \tag{3.171}$$

Let us take into account that the concentration of the ions (c_C, c_A) depends on the concentration of the salt (c), whereas $c_C = \nu_C c$, and $c_A = \nu_A c$. Introducing this, one can rearrange Eq. 3.171 and resolve it for $\Delta\psi$ in the following way:

$$\Delta\psi = -\frac{\mathrm{R}T}{\mathrm{F}} \left(\frac{z_C P_C \nu_C + z_A P_A \nu_A}{z_C^2 P_C \nu_C + z_A^2 P_A \nu_A} \right) \frac{\Delta c}{\bar{c}} \tag{3.172}$$

or:

$$\Delta\psi = \frac{\mathrm{R}T}{\mathrm{F}} \left(\frac{z_C P_C \nu_C + z_A P_A \nu_A}{z_C^2 P_C \nu_C + z_A^2 P_A \nu_A} \right) \ln \frac{c^{II}}{c^{I}} \tag{3.173}$$

The relation: $\Delta c/\bar{c} = \ln(c^I/c^{II})$ has already been introduced in Sect. 3.3.1 in context with Eq. 3.140.

It is easy to understand that Eq. 3.173 will be transformed into the Nernst Equation (Eq. 3.112), if the membrane becomes semipermeable, i.e., if $P_C = 0$, or if $P_A = 0$.

A better approach for the conditions of membranes will be the flux equation, derived by Goldman for conditions of constant electric field (see Sect. 3.3.2). Introducing this Goldman equation (Eq. 3.160) into the equation for electroneutrality of fluxes (Eq. 3.170), one gets the following expression:

$$P_C \frac{\mathrm{F}\Delta\psi}{\mathrm{R}T} \left(\frac{c_C^I - c_C^{II} \mathrm{e}^{\frac{\mathrm{F}\Delta\psi}{\mathrm{R}T}}}{1 - \mathrm{e}^{\frac{\mathrm{F}\Delta\psi}{\mathrm{R}T}}} \right) + P_A \frac{\mathrm{F}\Delta\psi}{\mathrm{R}T} \left(\frac{c_A^I - c_A^{II} \mathrm{e}^{-\frac{\mathrm{F}\Delta\psi}{\mathrm{R}T}}}{1 - \mathrm{e}^{-\frac{\mathrm{F}\Delta\psi}{\mathrm{R}T}}} \right) = 0 \tag{3.174}$$

This equation can also be rearranged and solved for $\Delta\psi$. For this we first transform the denominator of the fractions, using the expression:

$$1 - \mathrm{e}^{-x} = -\mathrm{e}^{-x}(1 - \mathrm{e}^{x})$$

This leads to:

$$\frac{\mathrm{F}\Delta\psi}{\mathrm{R}T \left(1 - \mathrm{e}^{\frac{\mathrm{F}\Delta\psi}{\mathrm{R}T}} \right)} \left[P_C \left(c_C^I - c_C^{II} \mathrm{e}^{\frac{\mathrm{F}\Delta\psi}{\mathrm{R}T}} \right) - P_A \, \mathrm{e}^{\frac{\mathrm{F}\Delta\psi}{\mathrm{R}T}} \left(c_A^I - c_A^{II} \mathrm{e}^{-\frac{\mathrm{F}\Delta\psi}{\mathrm{R}T}} \right) \right] = 0 \tag{3.175}$$

When $\Delta\psi \Rightarrow 0$ the expression in front of the square parentheses will not approach zero. Therefore, the sum inside the parentheses must be equal to zero:

$$P_C c_C^I - P_C c_C^{II} \mathrm{e}^{\frac{\mathrm{F}\Delta\psi}{\mathrm{R}T}} - P_A c_A^I \mathrm{e}^{\frac{\mathrm{F}\Delta\psi}{\mathrm{R}T}} + P_A c_A^{II} = 0 \tag{3.176}$$

Which gives after some rearrangements:

$$e^{\frac{F\Delta\psi}{RT}} = \frac{P_A c_A^{II} + P_C c_C^{I}}{P_A c_A^{I} + P_C c_C^{II}} \tag{3.177}$$

and:

$$\Delta\psi = \frac{RT}{F} \ln \frac{P_A c_A^{II} + P_C c_C^{I}}{P_A c_A^{I} + P_C c_C^{II}} \tag{3.178}$$

This is the *Goldman–Hodgkin–Katz equation* which is commonly used in electrophysiology to calculate diffusion potentials in living cells (mostly it is just named the *Goldman equation*). This expression also becomes a Nernst equation (Eq. 3.112), introducing the conditions of a semipermeable membrane ($P_A = 0$, or $P_C = 0$).

It is possible to extend this equation also for systems containing more than one salt, i.e., various monovalent ions. To take into account the nonideal character of solutions, chemical activities (a_i) instead of concentrations (c_i) can be used. In this case the *Goldman–Hodgkin–Katz equation* can be written as follows:

$$\Delta\psi = \frac{RT}{F} \ln \frac{\sum\limits_{\text{Anions}} P_A a_A^i + \sum\limits_{\text{Cations}} P_C a_C^e}{\sum\limits_{\text{Anions}} P_A a_A^e + \sum\limits_{\text{Cations}} P_C a_C^i} \tag{3.179}$$

When considering cells, the superscript i in this formula means "internal," the superscript e – "external" concentrations. Correspondingly: $\Delta\psi = \psi^i - \psi^e$.

In a large number of papers experiments are described indicating the applicability of the Goldman equation for various cells and membranes. Some limitations are obvious, however, when this equation is applied to real cellular conditions. Mostly these are already included in the conditions of the applied flux equation as discussed previously.

The most important limitation comes from the assumption considering a free diffusion of ions in a homogeneous and constant electric field. Even in the case of large pores in the membrane this assumption is valid only with some approximations. It must be taken into account that the thickness of a biological membrane is just as large as the Debye–Hückel radius of an ion. Additionally, the coefficients of permeability are defined for large systems and can be used for considerations of molecular systems only approximately.

As mentioned in Sect. 3.3.2, there exist a number of approaches that consider local concentrations of ions directly at the membrane boundary, to calculate fluxes and transmembrane potentials. This takes into account the influence of surface potentials and electric double layers (see Sect. 2.3.5). It is possible to introduce surface concentrations into the Goldman equation, instead of the concentration of ions in the bulk phase, using Eq. 2.10 or Eq. 3.113, if the surface potential ψ_0 is known. In this case, however, the value $\Delta\psi$ of Eq. 3.179 no longer means the

potential difference between the two bulk phases, as measured in electrophysiology by microelectrodes, but just between the inner and outer surface (see Fig. 2.48). Thus, this is an indication that the diffusion potential of a membrane is in fact controlled by local surface charges.

Further Reading

Goldman 1943; Katchalsky and Curran 1965; Syganov and von Klitzing 1999; Fraser and Huang 2007.

3.4 Membrane Transport and Membrane Potential

The living cell and its internal compartments maintain a particular electrolyte state which on the one hand guaranties constant conditions for all enzymatic processes, whilst on the other hand, it acts as an accumulator of electrochemical energy. This requires a complicated system of transporters in the membranes, which are particularly specialized and precisely controlled by a system of interactions. Furthermore, a number of receptor proteins transform external signals to internal information by modification of their transport properties. In a highly specialized way, nerve and muscle cells used the accumulated electrochemical energy for processes of excitation. In this chapter we will concentrate just on transport of ions with regard to transport of metabolites.

Numerous experimental investigations indicate an enormous variety of transporters even in a single cell. This research during the second half of the last century was supported strongly by the introduction of radioactive isotopes, later on by fluorometric methods and especially by the use of patch-clamp measurements. In the last decades, our knowledge on the molecular mechanisms of these transporters and their control has developed rapidly thanks to studies involving X-ray crystal analyses, and various high-resolution functional measurements.

3.4.1 Channels and Pumps: The Variety of Cellular Transport Mechanisms

Figure 3.25 illustrates a functional classification of various types of ion transport mechanisms in the membrane of cells and organelles. The differentiation between *pores* and *channels* is rather unclear. Mostly the term “pore” is used to denote larger membrane openings with low selectivity, which are for example produced by electric pulses (electric break down, see Sect. 3.5.5) or by other influences. In contrast “channels” are protein transporters, which are characterized by a certain selectivity. In any case, fluxes through these kinds of transporters are governed by the laws of electrodiffusion.

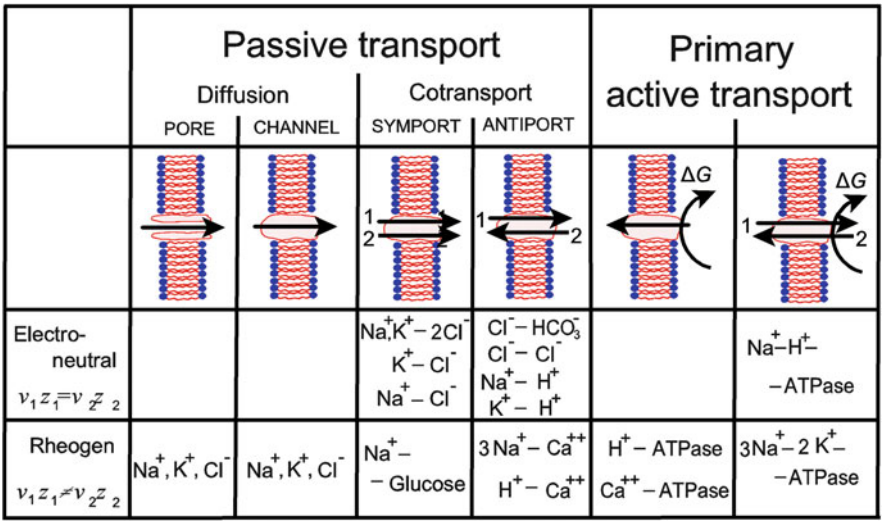

Fig. 3.25 Classification of various systems of ion transporters in biological membranes, including particular examples

Another type of passive transporters is so-called *carriers* or *porters* transporting simultaneously two or more ions in a well-defined stoichiometric relation. Such stoichiometrically coupled fluxes are called *co-transports*. There are two kinds of co-transport systems: In the case of the *symport,* a strongly coupled flux of two species in the same direction occurs. An example of this could be a complex that simultaneously transfers one Cl^- and one K^+ ion through the membrane in the same direction. In the same way, transport of an ion could be coupled to an uncharged molecule, like glucose. An *antiport*, in contrast to this, is a system simultaneously transporting two ions with identical charges in opposite directions, for example one K^+, against one H^+.

Co-transport systems are electroneutral, if an equal number of charges is transported, either of opposite sign in the case of a symport, or of the same sign in antiports. In this case the flux does not depend directly on electric field conditions. It is electrically silent, i.e., it cannot be identified by electrophysiological methods. In cases of unequal charge transporters, an electrical current will be the result of the transport. We will call this type of process *rheogenic*, i.e., "current producing." Rheogenic co-transport processes can be recognized by their electrical conductivity, a property which they have in common with simple diffusion processes. They can be controlled by electric fields, especially by the transmembrane potential.

An *active transport* is a sort of pump, transporting ions or even uncharged molecules against their own chemical or electrochemical gradients. Therefore, it is an "uphill transport," using metabolic energy (ΔG, in Fig. 3.25). In most cases these are so-called transport ATPases, using the energy of the hydrolytic reaction: ATP $\Rightarrow$ ADP. Furthermore, ionic pumps are known which are driven by other

sources of energy, such as for example decarboxylation, oxyreduction, or even the quantum energy of light. Some of these mechanisms can also run in the opposite direction. In chloroplasts and mitochondria, for example, ATP is synthesized by a "downhill" proton flux (see Sect. 4.8.3, Fig. 4.36).

Active transport can also be rheogenic. In this case the transport directly induces electric currents, like for example the Na-K-ATPase, transporting three charges from the inside out, but only two in the opposite direction, or a Ca^{++}-ATPase (see Sect. 3.5.2, Fig. 3.35). Frequently such transports are also called *electrogenic* which means: "generating an electrical membrane potential." Looking at the terms "rheogenic" and "electrogenic" accurately, they are however not identical. Even an electro-neutral pump can be "electrogenic" if it produces a concentration gradient of ions which subsequently generates a diffusion potential. Conversely, a rheogenic pump may influence the transmembrane potential only if there is a sufficiently high membrane resistance.

This leads to the differentiation between *primary* and *secondary* active transporters. An example of a primary active transporter is the Na-K-ATPase, where the uphill flux of ions is directly driven by a biochemical process. In contrast, secondary active transporters exploit the energy already stored in the electrochemical gradient of one species to drive the uphill transport of another substrate. This can be realized by various kinds of symporters or antiporters. As an example in Figs. 3.25 and 3.26 the co-transport of Na^+ with glucose is shown. It is "secondary active," because in fact the uphill glucose uptake is driven by the downhill Na^+-flux in a gradient, produced by the Na-K-ATPase. In a similar way fluxes of amino acids are coupled with transport of Na^+ or H^+ ions.

The number of different transport paths in a single membrane can be rather high. In Fig. 3.26 this is illustrated for the case of cells of a renal proximal tubule. It is

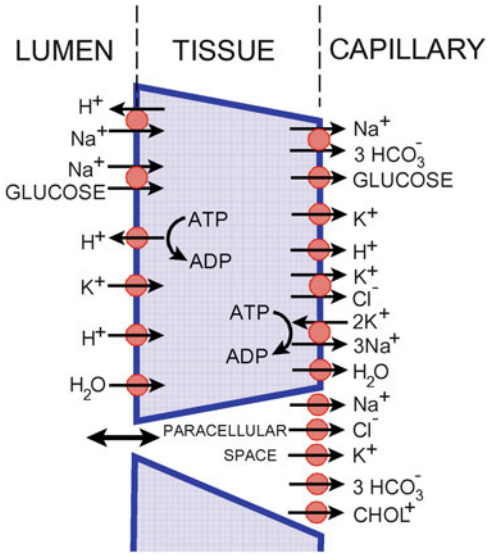

Fig. 3.26 Various types of ion transporters in cells of a renal proximal tubule and in the paracellular space (After Verkman and Alpern 1987)

obvious that the fluxes are coupled with each other by the transmembrane potential as well as by the concentrations of their common ions. Additionally, changes of fixed charges inside the cell induced by internal pH changes need to be taken into account.

The existence of transporters where the participants have strongly fixed stoichiometry forces us to rethink the requirement of flux electroneutrality which we postulated in Sect. 3.3.3 (Eq. 3.170). Considering rheogenic symports, it is not the electroneutrality of a single flux that is required, but rather the electroneutrality of all fluxes in the membrane of a single cell together. The calculation of the balance of charges and ions in a cell is therefore only possible by considering all fluxes. This type of coupling can formally be calculated using the flux matrix as discussed in Sect. 3.1.3.

The existence of co-transporters in a cell rather than simple diffusion processes can be regarded as a form of optimization. Ionic transport, based on electrodiffusion, strongly depends on the transmembrane potential. An alteration of the transmembrane potential would cause an immediate change of electrolyte fluxes in the whole cell, and subsequently a shift in the internal concentration of all ions. In contrast, the system of electroneutral co-transporters is independent of the transmembrane potential and will protect the cell against such disturbances.

Further Reading

Läuger 1991; Luckey 2008.

3.4.2 The Network of Cellular Transporters

If a cell were only a poly-electrolyte system without metabolically driven ion pumps it would remain in a state of Donnan equilibrium. This means that there would be a Donnan distribution of all mobile ions according to fixed charges, and as a result, a Donnan osmotic pressure (see Sects. 3.2.4, 3.2.5). In the living cell however, active transport systems driven by metabolic energy (Fig. 3.25) modify this ionic composition, as shown schematically in the model of Fig. 3.4b. The living cell therefore reaches a steady state, i.e., a stationary state of nonequilibrium in general (see Fig. 3.6), and for particular ionic species.

This nonequilibrium state has manifold functions. In general the cell can be regarded as a kind of electrochemical energy store which may be easily tapped. This, for example, is the case in electrical membrane de- and repolarizations (see Sect. 3.4.4). Furthermore, the nonequilibrium state of a system is the precondition for its homeostatic regulation. This, by the way, is also the reason for the increased temperature in homeothermic animals. The setting up of a concentration gradient of ions across the membrane makes the cells able to control and regulate an intracellular environment, which is the precondition of various cellular processes. In the case of Ca-ATPase an effective signal system is established. This pump creates an extremely low calcium level in the cytoplasm which is of the order of 10^4 times lower than the concentration in the extracellular fluid. In this way an important

signal transduction pathway is established, which can be triggered even by a minimal increase in the Ca-permeability of the membrane. The cytoplasmic Ca^{++}-ions act as a second messenger, in a number of cellular functions.

What therefore are the immediate effects of ionic pumps on the cell?

- They control and regulate the internal ionic milieu. In this way, steep gradients of the electrochemical potentials of particular ions are built up, essentially without changing the total internal ionic concentration. The internal potassium concentration of animal cells, for example, is usually much higher than the external one. Simultaneously however, the sodium concentration is lower to the same degree. The sum of both of these ions in the cytoplasm, taken together, is nearly the same as in the external medium.
- In the case of rheogenic pumps, they directly induce transmembrane potentials. In this case the pumps are called electrogenic.
- They can produce a direct osmotic effect changing the concentration of osmotically active substances.
- They can establish particular internal ionic conditions, controlling, for example, the extremely low intracellular calcium concentration.

Some direct consequences of the active transport processes can be demonstrated by the effects of stopping the pump through the use of specific inhibitors. In this case effects can be observed like Donnan-osmotic swelling, internal pH shifts, an increase in the internal calcium concentration, a change of transmembrane potential, etc. Mostly, using such inhibitors, the overall internal ionic conditions are altered.

As an example, the system of transport processes in kidney tubule cells is illustrated in Fig. 3.24. There are 13 different transport systems shown which determine the cellular milieu and additionally five other fluxes between the luminal and serosal surfaces of the epithelium across the paracellular gap. This picture in fact is incomplete as, for example, Ca^{++} fluxes are not shown, and the diagram does not include the intracellular organelles with their own transporters.

Using this example we will illustrate the interconnections of these transport properties qualitatively, following for example one particular path: Transport ATPases pump protons out of the cell, others decrease the internal sodium content, and in the same way enrich the cytoplasm with potassium. Extruding positive charges, both primary active transporters induce an inside negative transmembrane potential. Simultaneously, an electrochemical sodium gradient was generated which drives a sodium flux outside–in. This influx, however, is realized by a glucose-sodium co-transporter and acts therefore as a secondary active transporter for glucose entry. The glucose finally diffuses via its own concentration gradient on the opposite side of the cell layer from the cytoplasm into the capillary.

All these manifold transporters occurring in a single cell respond to different stimulants. Some of them become active only if a particular pH exists, others if the internal calcium concentration was increased. There are voltage-sensitive transporters responding to particular transmembrane potentials, or others that respond to mechanical stress of the membrane or to minimal temperature changes

(see Sect. 4.1). The electroneutral Na^+H^+ antiporter, which is present in most animal cells, merits particular attention. Under physiological conditions, at neutral pH_i it is inactive. However, if the internal pH increases, it becomes activated. This property qualifies it to be a volume-regulating system. This mechanism was demonstrated in the case of lymphocytes. It has also been shown that this Na^+ H^+ antiporter can be activated by a multitude of substances including hormones, growth factors, lectins, etc. These substances alter the above-mentioned pH threshold. This seems to be an important control mechanism for the regulation of complex biological phenomena.

Beside direct calculations of flux coupling, the equations of nonequilibrium thermodynamics can be applied to describe the energy balance of primary and secondary active transport. As an example the energy balance at steady state of the above-mentioned Na^+-Glucose symport will be evaluated. This is an example of a steady-state system like that of Fig. 3.4b which is determined by the active transport ($\mathbf{J}_A$) as well as by the passive flux ($\mathbf{J}_i$). In Sect. 3.1.4 we introduced the dissipation function $\mathbf{\Phi} = \sigma T$ (Eq. 3.64), which must be larger than 0. According to Eq. 3.64 for our system it amounts to

$$\Phi = \mathbf{J}_A\mathbf{X}_A + \mathbf{J}_i\mathbf{X}_i \quad \text{for:} \quad \Phi > 0 \tag{3.180}$$

In our particular case the glucose uptake ($\mathbf{J}_G$) is driven by the passive influx of sodium ($\mathbf{J}_{Na}$), driven by its electrochemical gradient. Corresponding to Eq. 3.180 this results in:

$$\mathbf{J}_G\mathbf{X}_G + \mathbf{J}_{Na}\mathbf{X}_{Na} > 0 \tag{3.181}$$

If v equivalents of sodium ions are transported for each mole of glucose then:

$$\mathbf{J}_G = v\mathbf{J}_{Na} \tag{3.182}$$

Introducing this in Eq. 3.181 and considering that both fluxes are not equal to zero, it follows that:

$$v\mathbf{X}_G + \mathbf{X}_{Na} > 0 \tag{3.183}$$

respectively:

$$\mathbf{X}_{Na} > -v\mathbf{X}_G \tag{3.184}$$

Let us now replace the forces ($\mathbf{X}$) by the differences of the corresponding chemical, resp. electrochemical potential (see Sect. 3.3.1), we obtain:

$$-v\Delta\mu_G < \Delta\tilde{\mu}_{Na} \tag{3.185}$$

Using Eqs. 3.33 and 3.41, and the conditions: $\Delta T = 0$ and $\Delta p = 0$, we get:

$$\nu \mathrm{R}T \ln \frac{a^i_{\mathrm{G}}}{a^e_{\mathrm{G}}} < -\left(\mathrm{R}T \ln \frac{a^i_{\mathrm{Na}}}{a^e_{\mathrm{Na}}} + \mathrm{F}\Delta\psi\right) \tag{3.186}$$

(where $\Delta\psi = \psi_i - \psi_e$) and after rearrangement:

$$\left(\frac{a^i_{\mathrm{G}}}{a^e_{\mathrm{G}}}\right)^{\nu} < \frac{a^e_{\mathrm{Na}}}{a^i_{\mathrm{Na}}} \quad \mathrm{e}^{-\frac{\mathrm{F}\Delta\psi}{\mathrm{R}T}} \tag{3.187}$$

This equation allows us to calculate the maximal rate of enrichment of glucose in the cell that can be achieved for a given electrochemical gradient of sodium ions. Assuming that the membrane potential of the cell is: $\Delta\psi = -50$ mV, and the relation of sodium ions: $a_{\mathrm{Na}}{}^{\mathrm{i}}/a_{\mathrm{Na}}{}^{\mathrm{e}} = 10$ ($T = 300$ K), it follows:

$$\left(\frac{a^i_{\mathrm{G}}}{a^e_{\mathrm{G}}}\right)^{\nu} < 69 \tag{3.188}$$

If the fluxes are coupled 1:1 ($\nu = 1$), this process gives a maximum enrichment of glucose by a factor of 69, when the pump is performing optimally.

Similar calculations can be applied to primary active transports, i.e., those that are driven by chemical reactions, for example transport ATPases. In this case in the equation of the dissipation function (Eq. 3.180), the reaction rate (as a type of scalar flux), and the chemical affinity of the energy supplying reaction (Eq. 3.75) must be included.

The calculation of the intensity of a pump which is necessary to build up a certain concentration gradient depends both on the coupling stoichiometry of the fluxes and on the passive back flow. This means that not only the power of the pump is responsible for the steady-state level achieved, but also the conductivity, resp. the permeability of the considered substance, leading it to flow backwards. This is illustrated in the scheme shown in Fig. 3.4b: the power of the pump must be higher if a greater difference in the levels of the vessels is reached, and if the outflow becomes faster.

Further Reading

Luckey 2008.

3.4.3 The Membrane Potential

As outlined in the previous section, the pumps lead to gradients of ion concentrations and therefore accumulate electrochemical energy. Now we will discuss how the cell generates an electrical membrane potential, using this accumulated energy.

First it is necessary to remember the general definition of electrical potential as defined in Sect. 2.2.1. According to this, the electrical potential [$\psi(x,y,z)$] is a scalar state parameter in three-dimensional space, similar to temperature (T) or pressure (p). Mostly as a simplification the function $\psi(x)$ is used to characterize the potential along a line that runs perpendicularly through the membrane (Figs. 2.15, 2.48). As the *transmembrane potential* ($\Delta\psi$) the potential difference is defined between two points, one on the inside, the other on the outside of the membrane, each at a suitable distance from it (Fig. 2.48). The sign of this difference results from its definition:

$$\Delta\psi = \psi_i - \psi_e \tag{3.189}$$

Note that terms such as Donnan potential, diffusion potential, Nernst potential, are just expressions describing the *mechanisms* which can give rise to the electrical transmembrane potential and do not refer in any way to different kinds of electrical potentials that might exist simultaneously. In fact there is only one electrical potential $\psi(x,y,z,t)$ at a given point in the space (x, y, z), and at a given time (t). In Fig. 2.48, the function $\psi(x)$ illustrates this in a very simplified way. It includes the transmembrane potential and the two surface potentials at both boundaries.

We have already learned that processes of active transport can be rheogenic (Fig. 3.25). If the so-far transported charges can be rapidly neutralized by other fluxes, for example by Cl^- exchange in the membrane of human erythrocytes, then a rheogenic pump has no direct electrical consequences for the cell. If however, no such short-circuit flux exists, the transported net charges build up a transmembrane potential, and the *rheogenic* pump becomes *electrogenic*.

In any case, the Na-K-ATPase, occurring in nearly all cell membranes, generates an electrochemical gradient of sodium and potassium. For most animal cells a relation near 1:10 occurs for $a^i_K > a^e_K$ and $a^i_{Na} < a^e_{Na}$. Chloride ions are distributed mostly passively, according to the Nernst equation. This nonequilibrium distribution of the cations can lead to a diffusion potential which can be calculated by the Goldman equation (Eq. 3.179) as follows:

$$\Delta\psi = \frac{RT}{F} \ln \frac{P_{Cl} a^i_{Cl} + P_K a^e_K + P_{Na} a^e_{Na}}{P_{Cl} a^e_{Cl} + P_K a^i_K + P_{Na} a^i_{Na}} \tag{3.190}$$

Even if the internal ion activities a^i_K and a^i_{Na} remain constant, the diffusion potential ($\Delta\psi$) can vary widely because of changing permeabilities (P_i). The limits of such variations can be easily obtained from Eq. 3.190:

For $P_K >> P_{Na}$, P_{Cl} Eq. 3.190 reduces to:

$$\Delta\psi_K = \frac{RT}{F} \ln \frac{a^e_K}{a^i_K} \tag{3.191}$$

and for $P_{Na} >> P_K$, P_{Cl} it follows:

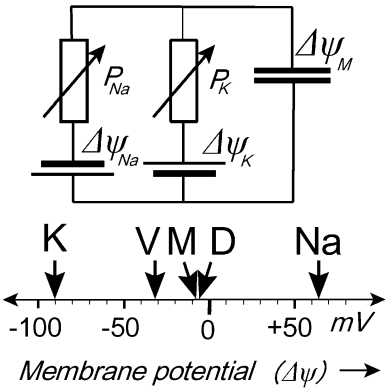

Fig. 3.27 An electrical circuit as a model illustrating the Na^+-K^+ diffusion potential of a cell as the result of a sodium ($\Delta\psi_{Na}$), and a potassium ($\Delta\psi_K$) battery. In the lower part of the figure, possible potential alterations are illustrated for the case of human erythrocytes in a solution containing 145 mM NaCl and 5 mM KCL. **K** – $\Delta\psi_K$, **V** – valinomycin-induced diffusion potential, **M** – potential of untreated erythrocytes corresponding to $\Delta\psi_M$, **D** – position of the Donnan potential, **Na** – $\Delta\psi_{Na}$

$$\Delta\psi_{Na} = \frac{RT}{F} \ln \frac{a^e_{Na}}{a^i_{Na}} \tag{3.192}$$

For these particular cases the Goldman equation (Eq. 3.190), therefore, reduces to a Nernst equation (Eq. 3.112) which was derived for such kinds of semipermeable membranes. If the typical relations of activities for sodium and potassium, as mentioned before, are inserted into Eqs. 3.191 and 3.192, then it is easy to understand that $\Delta\psi_K < 0$ and $\Delta\psi_{Na} > 0$.

This situation is illustrated in Fig. 3.27. The electrochemical gradients of potassium and sodium which are generated using metabolic energy can be considered as storage batteries, or electrical accumulators having opposite polarities. The permeability characteristics of the ions are expressed in this model as conductivities of the variable resistors, or potentiometers through which these accumulators are discharged. If the resistance is low, then a large discharge current would flow, and if the accumulator is not recharged continuously, it would soon be empty. In fact, the permeabilities P_{Na} and P_K are usually so low that the electrochemical gradient of the living cell persists for hours or even days. The effective membrane potential in this model is represented by the voltage difference across the capacitor $\Delta\psi_M$. This capacitor represents the capacity of the membrane (see Sect. 2.3.6). If P_{Na} and P_K have about the same value, then $\Delta\psi_M$ will be very small. If they differ, a membrane potential will be established according to Eqs. 3.191 and 3.192.

Figure 3.27 demonstrates membrane potentials that can be induced in human erythrocytes. In this case the Nernst potentials for potassium and sodium give the limits of these possible shifts. They range approximately between −95 mV and

+65 mV. The actual membrane potential of human erythrocytes in vivo is found to be −9 mV (**M**), and is only a little greater than the Donnan potential (**D**) which would result if the cell achieved a thermodynamic equilibrium (see Fig. 3.19). If the cells are treated with valinomycin, the membrane potential falls to about −35 mV (**V**). Valinomycin is an ionophore that is rapidly incorporated into the membrane causing a highly selective increase of potassium permeability. It will not reach the limiting value of the Nernst potential of potassium, because the values of P_{Cl} and P_{Na} are not negligible, as was assumed for Eq. 3.191. However, it is shifted in this direction.

Even if these types of potential alterations are possible without a significant change of concentration profiles, they must in fact be accompanied by a certain transmembrane shift of charges. It is easy to show that this charge flux is extremely small. For this we calculate the charge transfer across the membrane capacitor, which is required to adjust these potential differences ($\Delta\psi_M$ in Fig. 3.27). Let us ask the question: how many charges must be displaced in the cell membrane with a specific capacity of 10^{-2} F m^{-2} (see Sect. 2.3.6) in order to generate a transmembrane potential $\Delta\psi_M = 0.1$ V?

Equation 2.90 gives the corresponding relation for a capacitor. This enables us to calculate the surface charge density (σ) as a function of the transmembrane potential ($\Delta\psi$) and specific capacity (C_{sp}):

$$\sigma = C_{sp}\Delta\psi = 10^{-3}\text{C m}^{-2} \tag{3.193}$$

This value can be converted into charge equivalents of ions, using the Faraday constant (F):

$$\frac{\sigma}{\text{F}} = \frac{10^{-3}}{9.65 \cdot 10^4} \approx 10^{-8} \quad \text{charge equivalents} \cdot \text{m}^{-2}$$

The resulting charge density, so far, is very small. Considering a certain geometry of the cell, for example a sphere, or in the case of a neuron, a cylinder, one can easily transform this number into a concentration shift. The result will be a fully negligible proportion of the amount of internal ions.

This example demonstrates a most important element in the functional arrangement of the living cell: An ion pump driven by metabolic energy, accumulates electrochemical energy by generating a concentration gradient of sodium and potassium. This electrochemical energy can be converted into electrical energy altering the membrane permeabilities (for example: P_K and P_{Na}). In this way a wide-ranging control of the electric field in the cell membrane is possible. Even if the shift of the membrane potential amounts to only some tenths of a millivolt, the resulting variations of the field strength, sensed by the membrane proteins, are of the order of millions of volts per meter (see Sect. 2.2.1)! It must be emphasized that this control is possible without any sizeable input of energy and can be realized in milliseconds. Such permeability changes can be induced by the cell itself as well as by external influences.

As mentioned before, there are many ion-selective transporters in the cell which are controlled by internal calcium concentration, by internal pH, by mechanical tension of the membrane, or by modifications of other parameters. Diffusion potentials may also result from an interaction between the cell and specific drugs, or may be triggered locally through mechanical contacts with surfaces or particles, such as for example viruses. These alterations of membrane potentials caused by local permeability changes can induce electric potential differences and therefore electric fields not only in the x-direction, perpendicular to the membrane surface, but also in the y-, z-direction, i.e., in the plane of the membrane (see Sect. 3.5.2).

In the next section we will consider the action potential of nerve cells as a classical example of the feedback loop between an electric field and ionic permeability in more detail. Recently, the interest in the transmembrane potential of the cell as a regulator of cellular events has greatly increased. This concerns the size of the membrane potential in various cells, as well as its time dependence. Although action potentials have a special significance in signal transfer of neurons, they occur also in many other cells.

Although opening to particular transporters or integration of specific channels in the membrane may always modify the membrane potential by generating diffusion potentials, the resting potential of many cells is exclusively generated by electrogenic pumps. In this case transmembrane potentials appear to be independent of external potassium concentrations. Inhibition of the pumps in this case immediately leads to changes of $\Delta\psi$ (see Bashford and Pasternak 1986).

In Fig. 3.28 correlations of membrane potential and the state of various animal cells are illustrated. In contrast to cells with active proliferation like cancer cells or cells of embryos, indicating a transmembrane potential between -10 and -30 mV, nondividing cells, like neurons or skeletal muscle cells show membrane potentials between -70 and 90 mV. The transmembrane potential of cells which pass through a state of proliferation falls before mitosis takes place. It is not yet clear whether this reflects a regulatory mechanism of the cell, or whether it is only a phenomenon that accompanies such a mechanism.

In fact, in many cases alterations in the electrical field of a membrane seem to be of functional importance. The following mechanisms may cause this:

- The transverse component of an electrical field in the membrane may affect the functional state of intrinsic molecules. Dipole orientations for example, may modify the function of transport or other functional proteins, phase transitions in the lipid components of the membrane can be influenced by the field, or a transversal shift of small charged molecules can occur.
- The lateral component of the field can cause a displacement in its mosaic structure. This could lead to a local change in the mechanical properties of the membrane causing vesiculation, spike formation, etc.
- The electrical field can influence local ionic concentrations, as well as local pH values in close proximity to the membrane which, in turn, could affect transport processes, biochemical reactions at the membrane surface as well as receptor properties.

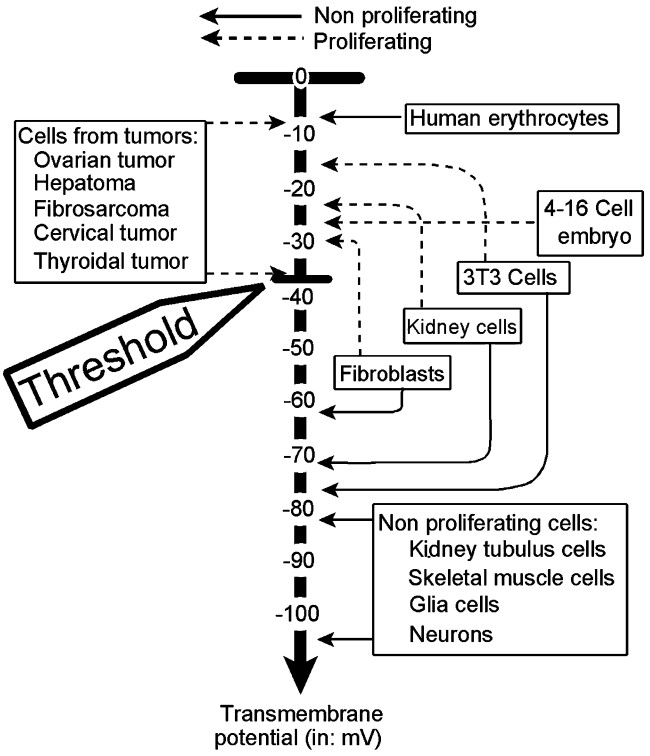

Fig. 3.28 The transmembrane potential of normal animal cells (*right*) and transformed tumor cells (*left*). It can be seen that proliferating cells indicate a membrane potential which is above the threshold value of −37 mV. Cells transiently arriving at the proliferating state lower their absolute potential. The human erythrocyte, as a non-nucleated cell with special physiological functions appears to be an exception (Drawn according to values from Bingeli and Weinstein 1986)

Further Reading

Glaser 1996; Starke-Peterkovic et al. 2005; Wang et al. 2003.

3.4.4 The Action Potential

In the previous section we described the possibility of cells to use the electrochemical gradient of potassium and sodium ions which is built up by active transport, to trigger various amounts of membrane potential simply by changing their permeabilities. This mechanism is expressed most efficiently in nerve and muscle cells. This was the reason why excitation phenomena were detected first in these cells.

Particular progress was achieved following the rediscovery of the giant axon of the squid in 1937 by John Zachary Young, and its subsequent introduction for biophysical measurements by Kenneth Stewart Cole. The use of these giant axons

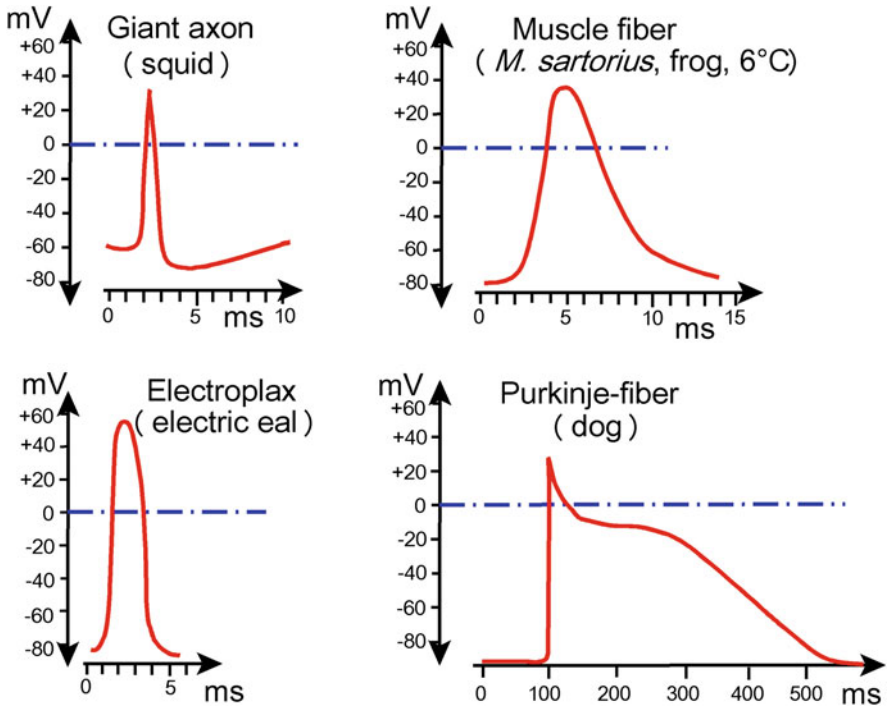

Fig. 3.29 Examples of various action potentials (After Penzlin 1991)

with a diameter up to 1 mm, have made it possible to apply the voltage-clamp technique to determine the ionic currents during the nerve impulse in extensive experiments by Alan Lloyd Hodgkin and Sir Andrew Fielding Huxley. In this technique, the electrical conductivity of the membrane is determined at various fixed transmembrane potentials, generated by microelectrodes. Recently, using patch-clamp techniques it has been possible to investigate the kinetics of these permeability alterations in extremely small membrane areas.

The action potentials of various nerve and muscle cells as illustrated in Fig. 3.29, can be qualitatively explained using the electrical scheme of Fig. 3.27 which was discussed in the previous section. The nonexcited nerve shows a very low sodium permeability (P_{Na}), its resting potential therefore, was determined chiefly by the diffusion potential of potassium which is negative inside-out. After excitation the membrane permeability for ions increased abruptly, whereas the sodium permeability rose quicker than that of potassium. For a short time therefore, the diffusion potential of sodium becomes dominant. This has the opposite polarity to the potassium potential which explains the spike of the action potentials.

As we demonstrated in the previous section the amount of charges that are needed for this kind of depolarization is extremely low. This was checked by flux measurements in excited nerves. During the generation of an action potential, therefore, no significant alterations of the internal ion concentration occur.

A nerve can generate action potentials for a long time after the ion pumps have been blocked. Only after hours does the electrochemical battery of the cell become empty.

Beside the entire mechanism of membrane excitation, the translation of the action potential along the axon of a nerve cell is of particular interest. In unmyelinated axons the process of pulse transmission is based on a lateral spreading of excitability by the electric field, generated by the excitation itself (see Fig. 3.37). The action potential generated by excited proteins in the membrane triggers the behavior of neighboring proteins. The impulse can proceed only in one direction, because of the refractory period of several milliseconds which the proteins need after an excitation to become excitable again.

Figure 3.30 illustrates the advantage of this kind of impulse propagation in relation to the transmission of a voltage pulse in an electric cable. In contrast to the cable, the time characteristics of the nerve pulse remains more or less constant, even after a certain distance of transmission. Conversely of course, the absolute velocity of pulse transmission in a cable is much faster than in an axon of a nerve.

The advantage of simple electrical conductivity is used in many vertebrate, and in a few invertebrate axons. In this case the axons are surrounded by *Schwann cells* forming the myelin sheath as an electrically isolating layer. Such nerves are called *myelinated*. This sheath is interrupted at intervals of some millimeters by so-called *nodes of Ranvier*, i.e., unmyelinated regions. In the myelinated regions simple electric conductivity of the pulse occurs, as in a cable. The nodes of Ranvier represent membrane areas which are excitable in a normal way. If a certain node of Ranvier is excited, then the pulse propagates by simple electric conduction along the myelinated length and excites the subsequent node. This so-called *saltatory conduction* is a form of pulse amplification leading to a faster transport of information. In contrast to about 1 m/s in unmyelinated nerves, the pulse propagation in fast myelinated nerves is up to 100 m/s.

In 1952 Hodgkin and Huxley, based on intensive experimental investigations on squid axons, proposed a theoretical model of membrane excitation in nerves (Nobel Prize 1963). Its form is of a purely kinetic nature and does not contain information about concrete molecular mechanisms taking place in the membrane.

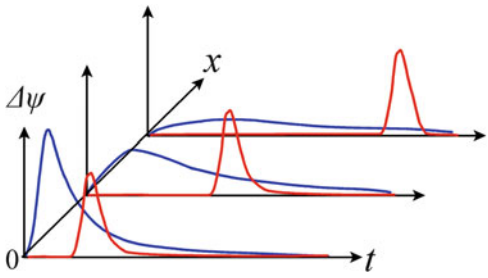

Fig. 3.30 The time course of a voltage pulse which is set at time $t = 0$ at point $x = 0$, transmitted in an isolated cable (*blue lines*) and in an unmyelinated nerve (*red lines*)

The basic equation describes the kinetics of the current in an electrical circuit, similar to the scheme in Fig. 3.27. The current density (j) in such a system can be described by the following equation:

$$j = C' \frac{\mathrm{d}(\Delta\psi_\mathrm{M})}{\mathrm{d}t} + (\Delta\psi_\mathrm{M} - \Delta\psi_\mathrm{K})G'_\mathrm{K} + (\Delta\psi_\mathrm{M} - \Delta\psi_\mathrm{Na})G'_\mathrm{Na} \tag{3.194}$$

$\Delta\psi_M$ is the electrical membrane potential, whereas the symbols $\Delta\psi_\mathrm{K}$ and $\Delta\psi_\mathrm{Na}$ indicate the Nernst potentials of potassium and sodium according to Eqs. 3.191 and 3.192. C' is the capacity of the membrane, and $G_\mathrm{K}{}'$ and $G_\mathrm{Na}{}'$ the potassium and sodium conductivities, always corresponding to a unit of area in the membrane. The conductivity of the membrane for individual ions cannot be measured electrically but can be obtained from experiments in which the kinetics of radioactive tracer ions is measured.

The first term of Eq. 3.194 gives the current density which leads to the charge of the membrane capacitor (Fig. 3.27). The following terms represent the current densities associated with potassium and sodium fluxes.

The conductivities $G_\mathrm{K}{}'$ and $G_\mathrm{Na}{}'$ are not constant, but functions of the electric field in the membrane, resp. of the membrane potential. The potentiometers in Fig. 3.27, therefore, are controlled directly by $\Delta\psi_M$. From the molecular point of view this means that these conductivities are the result of voltage-dependent channels. It is therefore necessary to make statements about field dependents of these conductivities, i.e., the functions $G_\mathrm{K}{}'(\Delta\psi_M)$ and $G_\mathrm{Na}{}'(\Delta\psi_M)$.

To describe the behavior of these channels, Hodgkin and Huxley used a statistical approach. They assumed that the channels can obtain only two discrete states: "open," or "closed." The phenomenological conductivities ($G_\mathrm{K}{}'$, $G_\mathrm{Na}{}'$) then represent the average of the functional states of a large number of such channels. If all of the channels are open then the maximal conductivities $G'_{\mathrm{K\ max}}$ and $G'_{\mathrm{Na\ max}}$ are established.

Furthermore, it is assumed that the potassium channel will be open when exactly four events take place simultaneously, all having the same probability of occurrence (n). The real nature of these events is not explained. It could be, for example, the presence of four potassium ions near the entrance of the channel.

This assumption leads to the following equation:

$$G'_\mathrm{K} = G_{\mathrm{K\ max}} n^4 \tag{3.195}$$

The probability n is a function of time and can be characterized by rate constants α_n and β_n as follows:

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \alpha_\mathrm{n}(1 - n) - \beta_\mathrm{n} n \tag{3.196}$$

Concerning the sodium permeability, it is assumed that the channel will be open when three events, each having the probability m occur simultaneously, and if

another inhibitory event having the probability h has not taken place. This leads to the expression

$$G'_{\mathrm{Na}} = G_{\mathrm{Na\,max}} m^3 h \tag{3.197}$$

For the parameters m and h also kinetic equations can be written:

$$\frac{\mathrm{d}m}{\mathrm{d}t} = \alpha_{\mathrm{m}}(1-m) - \beta_{\mathrm{m}} m \tag{3.198}$$

$$\frac{\mathrm{d}h}{\mathrm{d}t} = \alpha_{\mathrm{h}}(1-h) - \beta_{\mathrm{h}} h \tag{3.199}$$

The voltage dependence of the channels is proposed to be the result of influences on the rate constants α and β:

$$\alpha_{\mathrm{n}} = \frac{0{,}01(\Delta\psi+10)}{\mathrm{e}^{\frac{\Delta\psi+10}{10}}-1} \quad \beta_{\mathrm{n}} = 0{,}125\ \mathrm{e}^{\frac{\Delta\psi}{80}}$$

$$\alpha_{\mathrm{m}} = \frac{0{,}1(\Delta\psi+25)}{\mathrm{e}^{\frac{\Delta\psi+25}{10}}-1} \quad \beta_{\mathrm{m}} = 4\ \mathrm{e}^{\frac{\Delta\psi}{18}}$$

$$\alpha_{\mathrm{h}} = 0{,}7\mathrm{e}^{\frac{\Delta\psi}{20}} \quad \beta_{\mathrm{h}} = \frac{1}{\mathrm{e}^{\frac{\Delta\psi+30}{10}}+1} \tag{3.200}$$

(In these equations, the potentials are in mV!)

These equations were obtained from a purely empirical approach, analyzing measured parameters.

It is easy to see that if the relations given in Eq. 3.200 are substituted into Eqs. 3.196, 3.198, and 3.199, a system of nonlinear differential equations will be obtained. The solution of these equations can be substituted into Eqs. 3.195 and 3.197, and eventually, into the basic Eq. 3.194. An analytical solution of this system of differential equations is not possible. Computer simulations of these equations, however, indicate a good accordance with experimental results.

Figure 3.31 shows the calculated time courses for the changes in sodium and potassium conductivities at different membrane potentials. This also corresponds well with experimental findings. These curves illustrate the mechanism described above for the generation of an action potential. The conductivities from Fig. 3.31 illustrate the time-dependent changes of the potentiometers shown in Fig. 3.27, whereas the conductivities are directly proportional to the permeabilities. Within the first millisecond following the stimulus, the sodium potential is dominant because of the rapid increase in $G_{\mathrm{Na}}{}'$ (and thus P_{Na}). This will then be counteracted, by the increasing potassium potential.

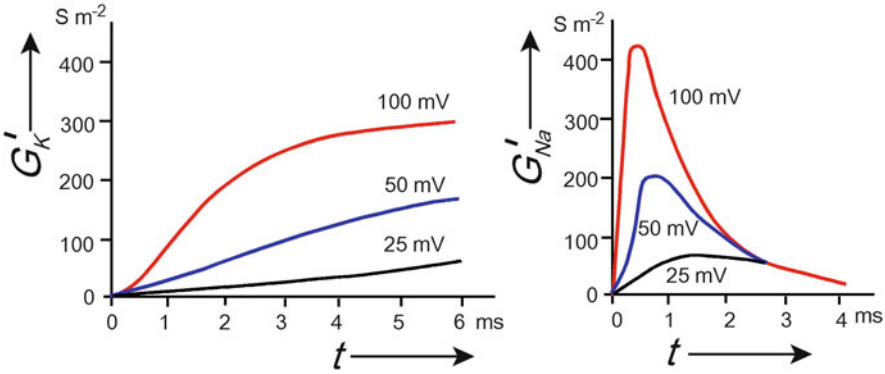

Fig. 3.31 The time dependence of the conductivities G_K' and G_{Na}' for various membrane potentials, corresponding to the theory of Hodgkin and Huxley

The Hodgkin–Huxley model and the corresponding measurements have provided a benchmark in our understanding of cellular excitability. New experimental techniques leading to more precise data nevertheless require some revisions of these approaches. So for example the mechanisms for the voltage-gated potassium and sodium ion currents have been superseded by more recent formulations that more accurately describe voltage-clamp measurements of these components. Especially its current–voltage relation has a nonlinear dependence upon driving force, corresponding to the Goldman–Hodgkin–Katz relation, rather than the linear approach used by Hodgkin and Huxley.

The original formulations of G'_{Na} and G'_K by Hodgkin and Huxley nevertheless continue to be used even though they do not adequately fit voltage-clamp measurements. The deviations between the m^3h and n^4 models (Eqs. 3.195 and 3.197), and the corresponding sodium and potassium currents do not appear to be eminently significant. Models that do describe these circumstances more precisely are more complex, which limits their practical utility in computational neuroscience.

Further Reading

Clay 2005; Hodgkin and Huxley 1952; Huxley 2002.

3.4.5 Molecular Aspects of Membrane Transport

In Sect. 3.4.1 various types of membrane transporters were characterized only in a phenomenological way. Now we will direct our attention to their structure and function. In fact, charged hydrophilic ions and molecules can penetrate the lipid membrane of cells and organelles only with the help of these mediators, usually proteins, the polypeptide chains of which span the lipid bilayer several times. In the last decades the molecular structure of a large number of these proteins has been revealed thanks to X-ray crystallography. In this way, the former more or less

mechanistic models of transport processes were replaced by more realistic molecular mechanisms.

In general, the following properties of transporters require an answer from these molecular considerations:

- Their extremely high selectivity including the phenomena of dehydration and rehydration of hydrophilic species in the process of membrane permeation.
- The mechanism of coupling between transport and the energy supporting biochemical reactions.
- The mechanisms of transport regulation by ligands and the transmembrane potential.

In 1998 MacKinnon unlocked the three-dimensional molecular structure of a potassium channel, a success which was awarded with the Nobel Prize in 2003. Such K^+ channels are found in bacterial as well as in eukaryotic cells of plants and animals, which are related members of a single protein family. Their amino acid sequences are easy to recognize because they contain a highly conserved segment called the K^+ channel signature sequence.

Let us answer some of the questions noted above using this extensively investigated example. The pore of this transporter is comprised of four identical subunits that encircle the central ion conduction pathway (two of them are depicted in Fig. 3.32). Each subunit contains two fully transmembrane α-helices, and a tilted pore helix that runs half way through the membrane. The hydrated K^+ ion, entering this channel from the cytoplasmatic side, first remains in the hydration state in a water-filled cavity with a diameter of 1 nm near the midpoint of the membrane. This cavity helps the K^+-ion to overcome the electrostatic repulsion that it would

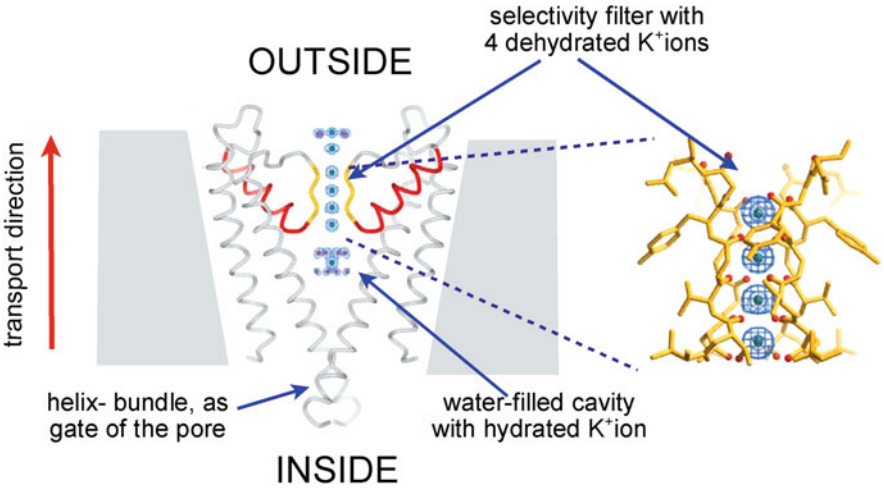

Fig. 3.32 The molecular structure of the KcsA channel. Only two subunits of this tetrameric molecule are shown. According to the position of the intracellular ends of the inner helices forming the gate, it is shown in a closed state (From MacKinnon 2003, modified)

normally experience when moving from the cytoplasmatic water phase into the low dielectric membrane environment. By allowing it to remain hydrated at the membrane center, and by directing the C-terminal negative ends of the protein helices toward the ion pathway, it becomes stabilized at the membrane interior. After this it enters the selectivity filter which contains four evenly spaced layers of carbonyl oxygen atoms, and a single layer of threonine hydroxyl oxygen atoms, which create four K^+ binding sites. In fact, on average only two K^+ ions are present at a given time in these four positions, always separated by one water molecule. It is very important that the arrangement of these protein oxygen atoms is very similar to that of water molecules around the hydrated K^+ ion. In this way the energetic cost of dehydration is minimized. Furthermore, a part of the binding energy is used for conformational changes of the proteins, which also is a prerequisite for the high conduction. In fact, the flux achieves up to 10^8 ions per second. This rate is large enough for sensitive amplifiers to record the electric current of a single channel. Na^+ ions cannot enter this filter because of their different crystal structure.

The gate of the channel is represented by a helix bundle near the intracellular membrane surface. In the closed position, as depicted in Fig. 3.32, the pore narrows to about 0.35 nm and is lined with hydrophobic amino acids, creating an effective barrier to the hydrophilic ions. This structure seems to be representative for many different potassium channels, irrespective of the stimulus that causes the pore to be in closed or open state. The conformational changes of these polypeptide chains that open and close the channel gate occur on the order of 10^2 times per second.

As discussed in previous chapters the membrane potential, and consequently the membrane electric field and its modification forms not only the basic principle of nerve and muscle excitation but regulates various functions in nearly all cells. This requires proteins, especially transporters, embedded in the membrane that sense alterations of this field and transform them into various cellular signals.

It is easy to imagine how an electric charge or an electric dipole can be reorientated within a protein when the field is changed. This can produce a conformational change in the protein that may regulate its function. The movement of the charge or the dipole induces a transient current (*gating current*) that can be measured experimentally and provides direct information about such conformational changes. The extent of the charge movement depends on the magnitude of the charge and the strength of the electric field in the region where the charge moves. In Sect. 2.2.1 (Fig. 2.15) as a crude estimation, this field strength was indicated to be of the order of 10^7 V m^{-1}. In fact, the exact value of this parameter near the corresponding charges or dipoles is unknown. In some cases the field can be concentrated to a narrow region around this location. Furthermore, the dielectric constant of this region inside the molecular structure is unknown.

The most extensively investigated voltage-gated channel is the so-called Shaker K^+ channel which can be expressed at a high density in *Xenopus* oocytes. It was isolated from *Drosophila melanogaster* and was named after the shaking that the fly

undergoes under anesthesia in its absence. Measurement of the gating current by patch-clamp techniques indicates that 13 electron equivalent charges per molecule are moving in this case. On the basis of the crystal structure of this protein, the so-called *paddle model* was proposed. It is assumed that voltage-gating charges are located on a hydrophobic helix-turn-helix structure, the so-called S4-segment, which can move within the membrane near the protein–lipid interface according to the direction of the electric field. Recently an S4-type sensor has been found in a voltage-dependent phosphatase, suggesting that this type of sensor may be modular and might have been incorporated into other types of proteins.

The kinetic model of nerve excitation as discussed in the previous section requires a particular sequence of opening and closing of potassium and sodium channels controlled by the membrane potential. Probably the four voltage-sensor domains of these channels react with individual time courses.

Although K^+ channels are excellent prototypes for voltage-gated channels, there are several other types of membrane proteins that differ in function, selectivity, regulation, kinetics, and voltage dependence. So for example a G-protein coupled muscarinic receptor has been found, in which a voltage-sensor is an integral part of the structure. It is expected that many other sensors will be discovered in the near future. More structures and biophysical analyses are still needed for a full molecular understanding of the function of these voltage sensors.

In contrast to the relatively simple mechanisms of channels, the pumps, and the co-transport systems require more functional elements, and the transport mechanisms demand more conformational changes in the corresponding transport protein. Especially the energy release by hydrolyzing ATP, and its coupling to ion movement needs a series of protein conformations. The first atomic-resolution structure of an ion pump was published in 2000 for the Ca-ATPase by Toyoshima et al. It shows an integral membrane protein with a large extended cytosolic part. In spite of the enormous progress of research in this field, a number of questions, especially concerning the Na-K-ATPase, are still open. The required conformational changes that accompany these transport processes mean that their speed is much slower than processes of channel transport.

The progress in determining the molecular structures of these channels has greatly facilitated the theoretical modeling and numerical simulation of the ion transport process itself. The most detailed description is based on the concept of *molecular dynamics* (MD). In this case microscopic forces of interactions between the penetrating ion and all atoms of the channel are calculated based on the classical Newton's equation of motion. This leads to trajectories of all the atoms in the system. In recent years, this approach has been used to simulate an increasing number of channels. Although this is the most detailed and accurate approach, it is limited by its shortcomings in quantitatively characterizing a large system, and its application depends considerably on advanced computational techniques.

Simpler and computationally less expensive of course are continuum models based on macroscopic or semimicroscopic continuum calculations like the Poisson–Nernst–Planck (PNP) approach. They, however, include a number of

limitations that have already been discussed in Sect. 3.3.2. A more realistic approach, situated between MD- and PNP-models, is based on *Brownian dynamics* (BD). In this case the trajectories of the ions in the system are followed using the Langevin equation. This is an approach, based on the theory of Brownian movement that considers the force not as the result of interactions of all atoms in the system, but rather of a subset of relevant coordinates. BD simulations have been applied to a variety of ion channels in recent years and the agreement with experimental work has been quite encouraging.

Further Reading

Bezanilla 2008; Faller 2008; Gadsby 2009; Kuyucak and Bastug 2003; Luckey 2008; MacKinnon 2003; Toyoshima et al. 2000; Zhou and Uesaka 2009.

3.5 Electric Fields in Cells and Organisms

3.5.1 The Electric Structure of the Living Organism

In Sect. 2.1.3 we pointed out that the term "structure" must be used in a very broad sense, not limiting it to visible details of a biological system. In fact, structures can be defined, based on the distribution pattern of various physical properties, such as concentrations, temperatures, streaming vectors, etc. Using the definition of the electric potential as a scalar state parameter $\psi(x, y, z, t)$ (Sect. 2.2.1), an *electric structure* can be established too. This, in fact exists on all levels of biological organization as a hierarchic structure, fully corresponding to the usual classification in biology, but governed by particular physical approaches (Fig. 3.33).

At the atomic and molecular level, the interactions can be explained using the approaches of wave mechanics. Particularly, the Schrödinger equation allows us to calculate the electric parameters at atomic dimensions, determining the energies of chemical bonds and molecular interactions.

Considering supramolecular structures like membranes, the electrical structure is determined by fixed and mobile charges and dipoles, forming electric double layers, and governing intermolecular interactions. Statistical thermodynamics can be used to consider these systems, leading to the Poisson–Boltzmann equation (Eq. 2.53). We already discussed electrical structures at this level of organization in Sects. 2.3.5 and 2.3.6.

For consideration of the cell as a thermodynamic system, the approaches of phenomenological thermodynamics were used. The interior and exterior medium of cells and organelles can be considered as phases with particular properties. Differences in the electrical potential between these phases, such as transmembrane potentials, can be described by the Nernst–Planck equation (Eqs. 3.156, 3.157). Their properties, dynamics, and biological relevance were discussed in Sects. 3.4.3 and 3.4.4. However, we remarked at that point that these properties cannot be

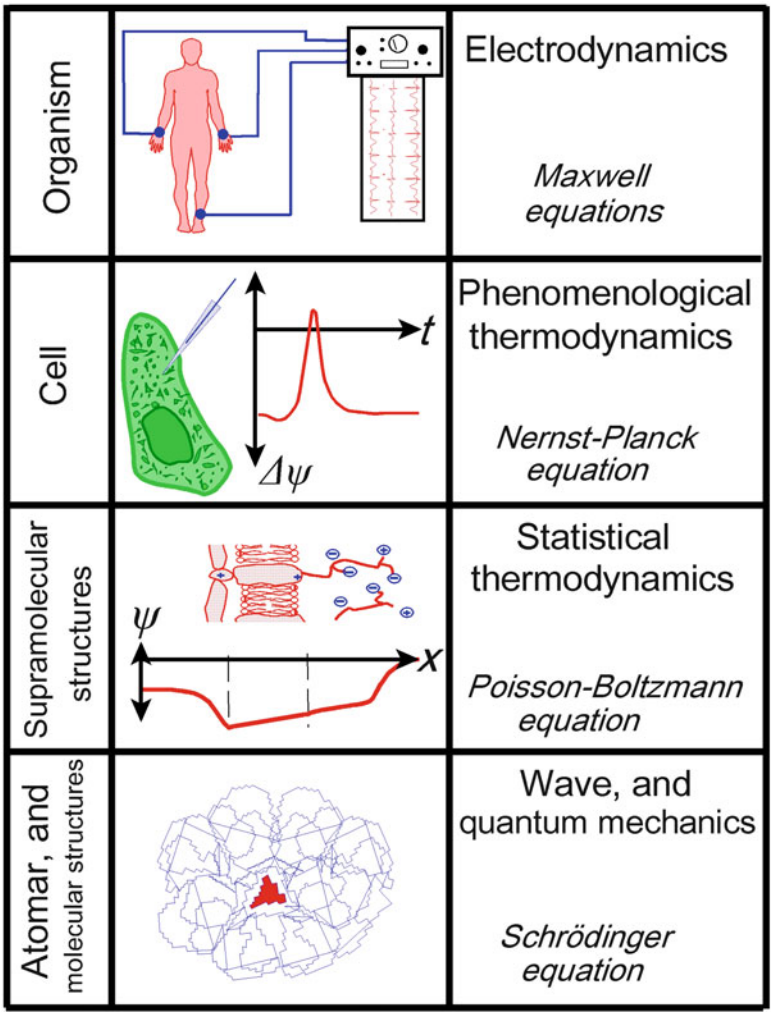

Fig. 3.33 The hierarchic system of the electric structure of the living organism and the corresponding physical approaches

calculated by phenomenological approaches alone. We mentioned the Poisson–Nernst–Planck theory (PNP-Theory, Sect. 3.3.2) as necessary for the completion of this approach.

In the following text we enter a further region of the hierarchic structure. We will consider electric fields in the extracellular space, in tissues and organs. This already extends into the area of classical electrodynamics where the Maxwell equations allow us to calculate electric fields in inhomogeneous dielectrics. The question arises: how does the field distribute inside the body, which consists of organs with different conductivities, like bones, soft tissue, air-filled cavities, etc.? We will

come back to this question again in Sects. 4.5, 4.6, and 4.7 where the influence of electromagnetic fields on biological systems is considered.

In spite of these hierarchic structure levels, it should be pointed out again that the electric potential is a physically defined parameter, independent of its source. Therefore, there exists only *one* electrical potential at *one* moment in time at *one* point in space. This statement is important with respect to the circumstances which will be discussed in the following text, namely the influence of externally applied electric fields on the intrinsic ones.

3.5.2 Extracellular Electric Fields and Currents

It has long been known that electric fields and currents exist not only across the membranes of cells and organelles but also in tissue and the whole body. Such fields are measured as electrocardiograms (ECG), electromyograms (EMG), and electroencephalograms (EEG) in medical diagnosis for example. EKG and EMG potential differences are of the order of several millivolts. EEG potentials are much lower because of the isolating role of the skull. Beside these oscillating potentials, it is also possible to detect DC potentials between various parts of the animal body and even in plants. Many measurements of these potentials, however, have suffered because of inaccurate use of electrodes.

The origin of extracellular electric fields and currents in biological organisms can be different. They can be generated directly as a result of ion transport in living cells or indirectly as streaming potentials or even by piezoelectric effects.

Let us first consider the electric conditions in the intercellular space (Fig. 3.34). As already mentioned (Sect. 2.3.6), there are strong electric fields around the fixed surface charges, not only perpendicular to the membrane plane, as considered in the electric double layer theory, but because of the lateral mosaic of surface charges also in the tangential direction. These fields, however, like the strong electric fields

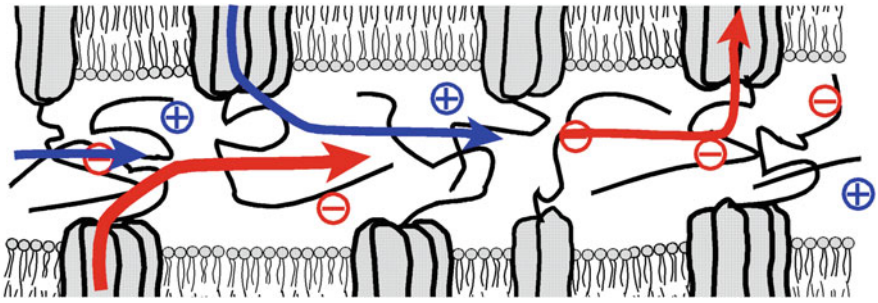

Fig. 3.34 The intercellular cleft can be considered as a Donnan system, formed by the mostly negative fixed charges of the surface coats of adjacent cells. This equilibrium, however, may be deflected by fluxes of positive and negative ions, extruded or taken in by transporters located in the membrane mosaic. Presently this picture remains speculative as there is no method to measure these parameters at the moment

in a double layer, are part of an electrochemical equilibrium, and do not generate electric currents. This Donnan system, however, is deflected by ion transport processes through the adjacent cell membranes. As illustrated schematically in Fig. 3.26, a large variety of rheogenic transporters permanently move ions across the membrane at various locations. This results in a complicated current pattern through the system of fixed charges around the cells, and in the intercellular clefts. Unfortunately, this situation is just speculative because to date no method exists to analyze this situation experimentally in intercellular clefts, or in the dimensions of surface coats of cells.

In some cases the distances of the individual electrogenic transporters are large enough to produce currents which are measurable in their vicinity. One example is a rheogenic calcium pump which is localized at a particular point in the membrane of eggs of the fucoid seaweed *Pelvetia*. This pump induces a current density of up to 0.3 A m^{-2}, and an electric field around the cell. It is possible to measure this field using the vibrating probe technique (see Fig. 3.35). For this a glass microelectrode is used, the top of which consists of a metallic sphere with a diameter of a few micrometers. This electrode vibrates at a frequency of several hundreds of Hz causing this sphere to be displaced periodically at an amplitude of 10–30 μm. Using a very sensitive low noise amplifier it is possible to measure voltage differences near 1 nV = 10^{-9} V between these two reversal points.

Interestingly, this field seems to play a role in growth regulation of these eggs, and determines the direction in which the first rhizoid of the cell will grow. Experiments with external fields confirmed that in fact, the field determines the direction of growth.

Meanwhile, a large number of biological objects have been investigated with this method, such as single cells, as well as various growing parts of plants, follicles of insect eggs, growing embryos, muscle fibers, and other tissues and organs. Investigation of embryos of the clawed frog *Xenopus* showed particular currents

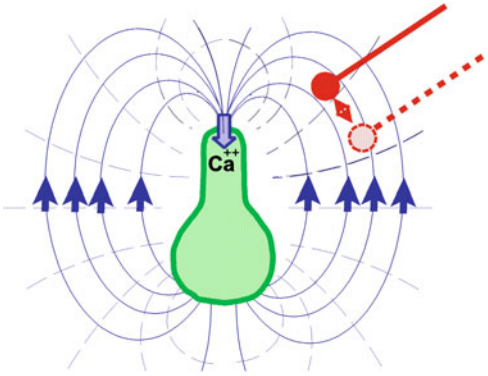

Fig. 3.35 Current- (——), and equipotential lines (- - - -) near a growing egg of the brown algae *Pelvecia*, caused by local rheogenic calcium transport, which determines the location where the rhizoid is formed. The external field was measured using the vibrating probe method. The corresponding electrode is shown in two extreme positions during vibration. Relations of electrode and cell sizes are approximately correct (According to data from Jaffe 1979)

and fields that change direction and intensity at different stages of development. Obviously, this reflects a particular control function, the mechanism of which is unclear. It is possible that in this way the diffusion of charged growth factors is directed from one group of cells to another. In this case the vector orientation of the electric field is imposed on the chemical gradient.

A special source of electric currents and fields are the so-called *wound potentials*. The reason for them is permanent electrostatic potential differences between different cavities in the body, caused for example by ion pumps, differences in electrolyte composition and others. This leads to various transepithelial potentials. An intact mammalian corneal epithelium for example maintains a potential difference of about 40 mV. It results from net inward transport of K^+ and Na^+, and the net outward transport of Cl^- to the tear fluid. In mammalian skin the inward transport of Na^+ leads to a potential difference of about 70 mV between dermis and the outer layer of the epidermis. The maintained potential difference is possible due to tight junctions, which are closely associated areas of the cells forming a barrier with high electrical resistance.

If the isolating properties of an epithelium are disturbed by a wound, the potential collapses at this point because short circuit currents occur. This leads to an electric field with a lateral orientation to the plane of the epithelium (see Fig. 3.36). In most cases the wound has a positive polarity in relation to the surrounding surface. As a noninvasive method the above-mentioned vibrating electrodes are also used for mapping the electric field near wounds. For this a small metal vibrating probe with a displacement of 0.18 mm in air above the skin measures the surface potential of the epidermis through capacitive coupling. In Sect. 4.4.2 we will indicate how cells may use this field by way of galvanotaxis to close a wound.

Another source of extracellular currents and fields are various membrane excitations. The axon of a nerve cell may be the best example. If an action potential propagates along the axon, small areas of this membrane will become depolarized. In contrast to the parts of the axon with resting potential, where the external membrane side is positively charged in relation to the inner one, the polarity of membrane sides with action potential is opposite. The result is a lateral electrical current from one point of the membrane to another (Fig. 3.37). Such local depolarizations not only occur in nerves, but also in muscle and other cells.

In the case of a synchronized excitation of a bundle of nerve or muscle cells, the currents and fields of the single cells are superimposed, and can be measured as electrocardiograms (ECG), electromyograms (EMG), and electroencephalograms (EEG) on the periphery of the body. The use of surface electric potential differences in medical diagnosis raises the question: how do electric fields, generated for example by the beating heart, spread in the body? Modern techniques allow a reasonably good reconstruction of the electric field distribution, using parallel recordings of the ECG simultaneously in different parts of the body. The measurable ECG results from excitation of definite parts of the cardiac muscle in the rhythm of the beating heart. Consequently, the heart becomes an oscillating dipole, the orientation of which changes according to the orientation of the actually excited parts of the muscle.

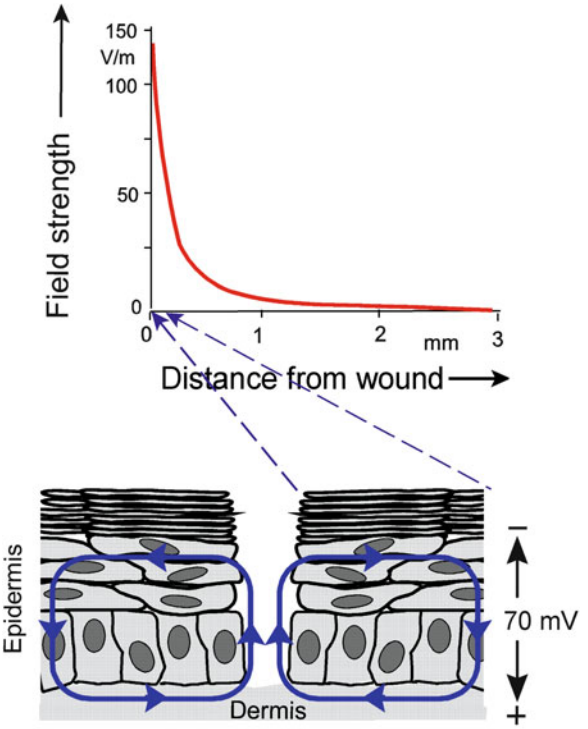

Fig. 3.36 Schematic representation of the occurrence of a wound potential in the mammalian epidermis. The field strength near the wound edge reaches 140 V m^{-1} and declines strongly with distance (Data from McCaig et al. 2005)

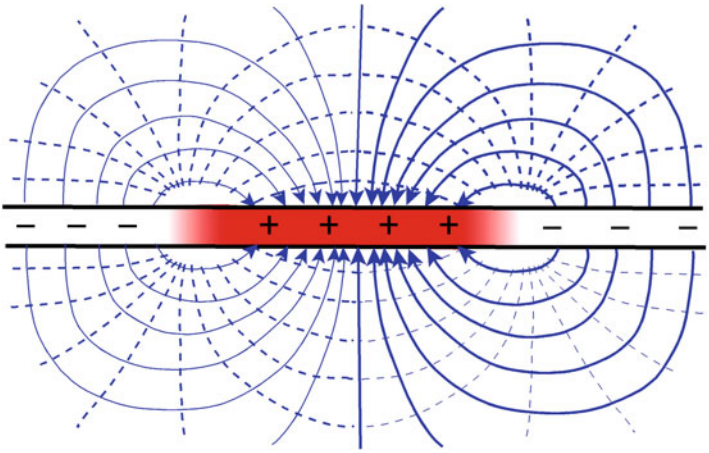

Fig. 3.37 Schematic illustration of a snapshot of an excited nerve. The *red areas* represent the actual location of the depolarized membrane (see also Fig. 3.29)

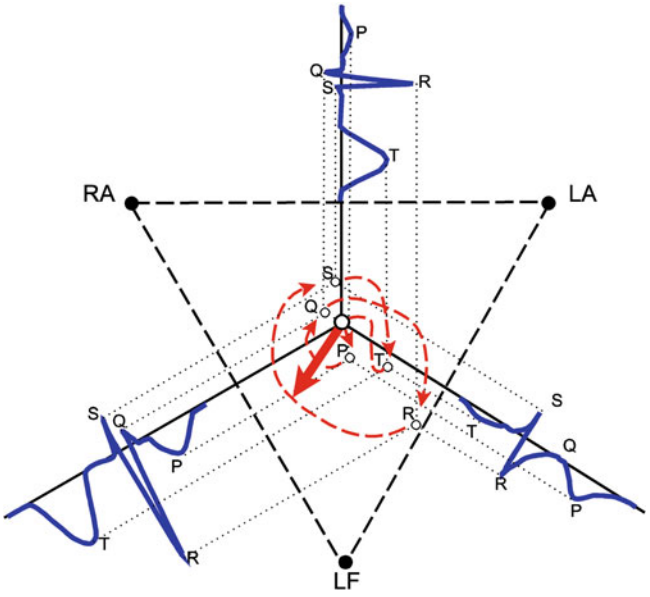

Fig. 3.38 Construction of a vector cardiogram according to the Einthoven triangle. Using the time course of the potential curves (*blue*), orientated in an equilateral triangle, a rotating dipole (*red arrow*) can be constructed in the center. P, Q, R, S, T – are the corresponding waves of the ECG. *RA* right arm, *LA* left arm, and *LF* left foot

The first person to propose a method to evaluate ECGs was the Dutch physiologist Willem Einthoven, who was awarded the Nobel Prize for his work in 1924. He proposed that it should be possible to localize the excited parts of the heart by detecting the potentials at three points on the body, which are more or less equidistant from the heart. This so-called *Einthoven triangle* is illustrated schematically in Fig. 3.38. The three cardiograms, derived from the corresponding points, indicate a periodic sequence of P, Q, R, S, and T waves. These waves represent the sequence of excitation of different parts of the heart muscle, starting with the excitation of the atria (P wave). If all the atrial fibers are in the plateau phase, the PQ segment is reached. Subsequently, the excitation spreads over the ventricles, beginning on the left side of the ventricular septum, spreading toward the apex (QRS waves), and finally reaching the ventricular recovery phase (T wave). As a result of the projection of these curves corresponding to the geometry of an equilateral triangle, a rotating vector appears, the origin of which lies in the crossing point of the three axes. The arrowhead moves periodically along the dashed line. Because of the dielectric heterogeneity of the body, an exact correlation between the resulting dipole vectors of the field with the anatomical position of various parts of the heart muscle is impossible.

In a special adaption, electric fishes may use extracellular fields generated by excitable cells. In the electric organs specialized muscle cells are organized in so-called *electroplaques*, where the voltage of several cells adds up to

approximately 800 V. Such high voltage is only possible, however, in freshwater fishes like the electric eels. Marine fishes generate smaller voltages because of the high conductivity of seawater. In the electric organs of these fishes the cellular elements are not arranged in series, leading to higher voltages, but rather in parallel to increase the current. So-called *strong electric fishes* use the induced electric fields to catch their prey, or to defend themselves, whilst *weak electric fishes* use electric fields only for orientation. Meanwhile, the amazing capacity of this sensory system has been determined. Although it works only over small distances, which amounts approximately to half of the body length of the fish, shapes and dielectric properties of subjects in muddy water are detected in this way (for electroreception, see Sect. 4.5, Fig. 4.25).

Beside these kinds of currents and fields generated by living cells, other sources in the body occur as a result of electromechanical transformations, for example by piezoelectric and electrokinetic effects (Sect. 2.3.5). Both effects occur in bones and cartilage during in vivo deformations.

Piezoelectricity is comprised of translocations of charges in crystals or crystalloid macromolecular structures resulting from mechanical loading. In bones piezoelectric potentials are mostly the result of deformations of collagens. Furthermore, under mechanical stress, such deformations cause a flow in the narrow channels of the bone with negative surface charges. This additionally leads to streaming potentials. In contrast to the electric fields in the whole body induced by nerves and muscles, these fields are rather low at larger distances. Conversely, they seem to be important locally in the process of bone growth and bone remodeling in vivo. This circumstance has been used in efforts to stimulate bone repair by applying electric field pulses or ultrasound.

Further Reading

Electric fields near cells: McCaig et al. 2005; Nuccitelli et al. 2008; Shi and Borgens 1995; EEG: Nunez and Srinivasan 2006; Reilly 1998; electric fields in bone: MacGinitie 1995; electric fishes: Fortune 2006; Peters et al. 2007.

3.5.3 Passive Electrical Properties of Tissue and Cell Suspensions

In contrast to the membrane potential and the electrical currents driven by rheogenic pumps, all requiring *active* biological processes, other qualities like electrical resistance, conductivity, or membrane capacity can be summarized as *passive electric properties* of biological systems. These parameters are essential not only to calculate the distribution of the actively generated fields and currents in the organism, as described in the previous section, but also the penetration and induction of currents in the body caused by external influences (Sect. 4.6.1). This includes AC fields over a broad frequency range.

To derive the basic parameters and equations, let us first consider the electrical properties of a plate capacitor filled with material of certain conductivity and

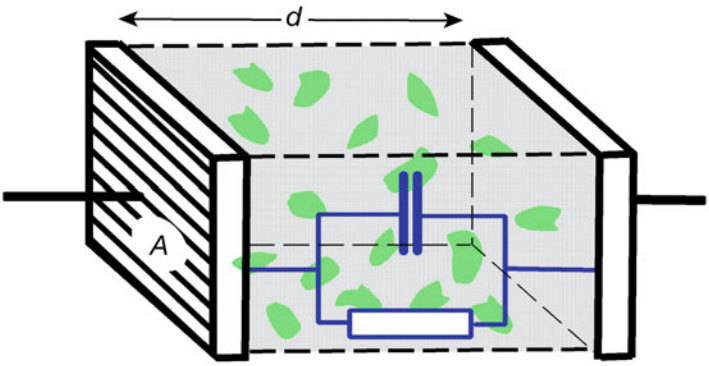

Fig. 3.39 A capacitor consisting of two parallel plates of area A, and a mutual distance d, filled with an inhomogeneous medium (e.g., cell suspension or biological tissue). Neglecting the real dielectric heterogeneity, it can be described formally by a RC-circuit (*blue*)

dielectric constant, such as for example biological tissue. Irrespective of its dielectrical heterogeneity, this system formally can be described by an equivalent RC circuit consisting of a conventional capacitor and a resistor in parallel (Fig. 3.39).

An AC voltage applied to such a system generates a current which will flow through the resistor and periodically recharge the capacitor. This results in an effective AC resistance which is called *impedance*, whereas its reciprocal is the AC conductance or *admittance* (Y^*). In fact, this admittance is determined by the static conductance (G) of the resistor, and the frequency (ω) dependent displacement current passing through capacitor. This behavior can be summarized as follows:

$$Y^* = G + j\omega C \tag{3.201}$$

This equation includes the imaginary number $j = \sqrt{-1}$ transforming the admittance (Y^*) into a complex parameter, marked by the superscript *. The reason for this is the response of the system to the particular time function of the AC-current. For the behavior of an RC circuit in an AC field not only the amplitude of a sine current must be taken into account, but also the occurring phase shift which is coded by the term j.

To understand this, let us consider an AC voltage having the following time function:

$$U = U_{\max} \sin \omega t \tag{3.202}$$

An applied voltage (U) therefore oscillates with an angular frequency ($\omega = 2\pi v$) and a peak amplitude of $U_{\max}$. According to Ohm's law (Eq. 3.51), for a circuit containing only a resistance the current (I) possesses the same time behavior:

$$I = I_{\max} \sin \omega t \tag{3.203}$$

In an RC circuit, however, additionally the displacement current of the capacitor must be taken into account. The charge (q) of a capacitor with the capacity (C) is determined by the following equation:

$$q = U\ C \tag{3.204}$$

Since the current is defined as the time derivative of charge, for the time-independent capacity (C), the displacement current (I_C) is:

$$I_C = \frac{\mathrm{d}q}{\mathrm{d}t} = \frac{\mathrm{d}(U\ C)}{\mathrm{d}t} = C\frac{\mathrm{d}U}{\mathrm{d}t} \tag{3.205}$$

Introducing Eq. 3.202 into Eq. 3.205 for C = const, one gets:

$$I_C = C\frac{\mathrm{d}(U_{\max} \sin\omega t)}{\mathrm{d}t} = C\ U_{\max}\omega \cos \omega t = C\ U_{\max}\omega \sin\left(\frac{\pi}{2} + \omega t\right) \tag{3.206}$$

Further, defining:

$$I_{C\ \max} = C\ U_{\max}\omega \tag{3.207}$$

it follows:

$$I_C = I_{C\ \max}\sin\left(\frac{\pi}{2} + \omega t\right) \tag{3.208}$$

Comparison of Eq. 3.202 and 3.208 clearly indicates for a simple capacitor a phase shift of $\pi/2$ in between current and voltage.

Consequently, for the AC current in Eq. 3.201 this phase shift, for example the time course of the capacitive component of the sum, must be considered. This circumstance is taken into account using a Gaussian plane containing imaginary numbers. For this in Eq. 3.201 the term ωC is multiplied by the imaginary number $j = \sqrt{-1}$ and plotted on the ordinate in these Gauss coordinates.

Equation 3.201 can be modified using the geometrical parameters of the electrode: area (A) and mutual distance (d), together with the material constants: specific conductivity (g), and permittivity ($\varepsilon\varepsilon_0$), whereas:

$$G = \tfrac{A}{d}g \quad \text{and} \quad C = \tfrac{A}{d}\varepsilon\varepsilon_0 \tag{3.209}$$

Introducing these relations into Eq. 3.201, one gets:

$$g* = g + j\ \varepsilon\ \varepsilon_0\omega \qquad \text{whereas :} \qquad Y^* = \frac{A}{d}g^* \tag{3.210}$$

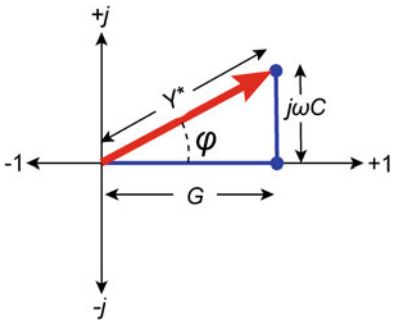

Fig. 3.40 Representation of the conductance (G) of the resistor and the admittance of the capacitor ($j\omega C$) of the analog circuit from Fig. 3.39 in the Gaussian plane of complex numbers. The admittance of the system (Y^*) corresponds to the length of the resulting vector. The angle φ represents the resulting phase shift

In this equation g^* is called the *complex specific admittance*.

In the same way as the complex specific admittance (g^*), a complex dielectric constant (or *relative permittivity*) (ε^*) of the system can be formulated. Starting with the equation for the effective AC resistance of a capacitor ($R = 1/\omega C$), one can introduce a complex capacitance (C^*) as follows:

$$C^* = \frac{Y^*}{j\omega} \tag{3.211}$$

Inserting the parameters of Eqs. 3.201 and 3.209 into this equation, and considering $C^* = \varepsilon^* \varepsilon_0 A/d$, and $1/j = -j$, one gets:

$$\varepsilon^* = \varepsilon - j\frac{g}{\omega \varepsilon_0} \tag{3.212}$$

The derivation of these basic equations (Eqs. 3.210 and 3.212) allows us to understand the properties of complex dielectrics in AC fields.

In fact, the electrical properties of biological tissues containing a multitude of various cells with various size and intercellular clefts are too complicated to be described by a simple RC circuit as shown in Fig. 3.39. An enormous network of elementary RC elements with different parameters would be required to build an appropriate model. This, of course, is impossible for practical use. For this reason one uses the simplified scheme as shown in Fig. 3.39 with the knowledge that the properties of the resistor as well as the capacitor are now frequency dependent.

Figure 3.41 shows mean values of the complex dielectric constants (ε^*) and the specific admittance (g^*) of various tissues over a broad frequency range. It can be seen that the specific admittance increases with frequency. The main reason for this is the increase of the membrane admittance. Conversely, the dielectric constants are

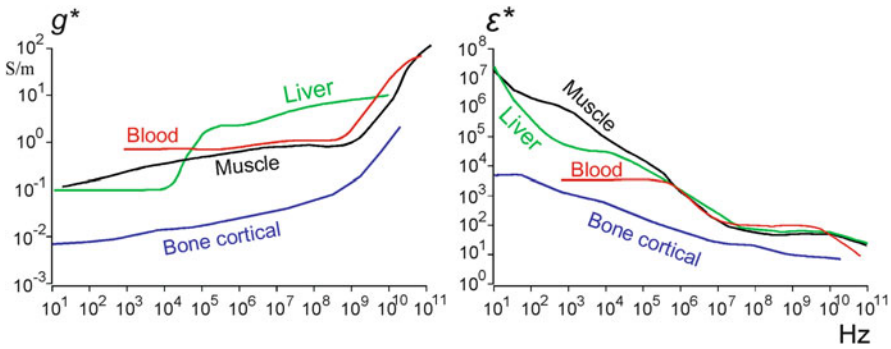

Fig. 3.41 Complex specific admittances and complex dielectric constants of various tissues as functions of frequency (According to averaged data from Gabriel et al. 1996)

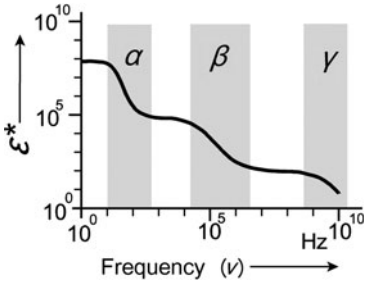

Fig. 3.42 Schematic illustration of the frequency regions of the α-, β-, and γ-dispersions of the complex dielectric constant of a biological tissue

extremely high for tissues at frequencies below the MHz range. They drop to the standard value of the dielectric constant of water ($\varepsilon \approx 80$) only at microwave frequencies.

Typically, the dielectric constant of tissues or cell suspensions decreases in a characteristic step-wise fashion. The reasons for this are the properties of various kinds of RC-circuits in the system, each with different time constants of relaxation. These frequency regions, in the first instance have formally been designated as α-, β-, and γ- dispersions (see Fig. 3.42).

Various phenomena are summarized as *α-dispersion*, which for cell-sized objects occurs in the frequency range below 10 kHz. Mostly, reactions in the electric double layer of the cell membranes are responsible for this, such as various electrokinetic phenomena (see Sect. 2.3.5), for example the deformation of ionic clouds. Because of various artefacts, related to polarization phenomena of the electrodes, and electro-osmotically induced convections, it is difficult to measure the real physical parameters in this frequency region. It is difficult to determine the individual components of an inhomogeneous dielectric system.

The *β-dispersion*, also designated as *Maxwell–Wagner dispersion*, is mostly based on processes connected to membranes as dielectric barriers. Consequently these dispersions are caused by structural inhomogeneities of the material, such as

cellular and organelle structures. Some authors subdivide the β-region into β_1-, and β_2-ranges which refer to the dispersion of the cell membrane (β_1), and cytoplasm (β_2) polarization dispersions, respectively.

The *γ-dispersion* at higher frequencies is caused by so-called *Debye relaxations* of various molecular dipoles. At frequencies of the γ-dispersion region, even the resistivity of the internal and external milieu of the cell cannot simply be described by ohmic resistances (see the small RC-circuits in the resistors of Fig. 3.43).

The dispersion of water dipoles occurs at 18.7 GHz. This in fact, is true only for free water. Recent measurements indicate that the dispersion of bound water may occur at frequency regions even below 1 GHz. This is important because of possible effects of high-frequency electromagnetic fields on biological tissue (Sect. 4.7.1).

The measurement of dielectric properties of solutions and heterogeneous media, the so-called *impedance spectroscopy* is used in various applications. Furthermore, electrical impedance tomography (EIT) has been developed for imaging of particular dielectric properties of parts of the body. For measurements a multitude of electrodes is placed on the skin. Proposed applications include the monitoring of lung function, detection of cancer in the skin or breast, and the location of epileptic foci. This method, however, is still at an experimental stage and not used in routine diagnostics.

Further Reading

Barnes and Greenbaum 2006, Gabriel et al. 1996; Holder 2005; Orazem and Tribollet 2008; Pethig and Kell 1987; Riu et al. 1999; Schwan 1957.

3.5.4 Single Cells in External Electric Fields

Cell suspensions or single cells are exposed to electric fields of various frequencies in many biotechnological approaches (see Sect. 3.5.5). Therefore, it is important to analyze field distribution and currents through and around these cells.

Figure 3.43 indicates a simplified electronic circuit, describing the passive electric properties of a spherical cell in an electrolyte solution. As in Fig. 3.27, the membrane is represented by a capacitor (C_m) in parallel with a resistor (R_m) simulating the membrane resistance. At DC, and low-frequency AC fields, the conductivities of the external and internal media can be described by simple resistors (R_{e1}, R_{e2}, R_i).

The high specific conductivity of the cytoplasm and the external medium on the one hand, and the extremely low conductivity of the membrane on the other results in R_m being more than 7 orders of magnitude higher than R_e or R_i. Applying Kirchhoff's law, a low frequency current therefore does not flow through the cell, but around it.

This is demonstrated schematically in Fig. 3.44 where a spherical cell is shown with the electrical properties as described above. In case of DC fields and extremely

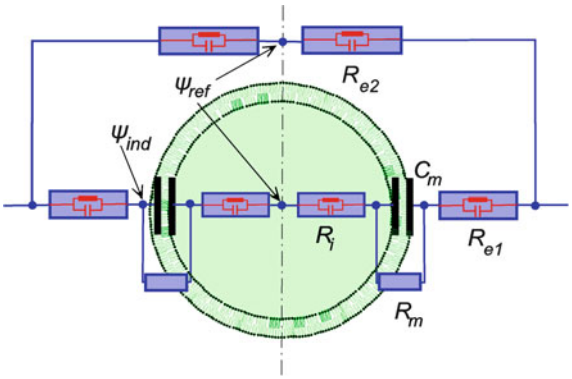

Fig. 3.43 Simplified analog circuit demonstrating current paths through and around a spherical cell. R_{e1} and R_{e2} – resistances of the external medium, R_i – resistance of the cytoplasm, R_m – membrane resistance, C_m – membrane capacity, ψ_{ref} – reference potential at the symmetry plane of the system, ψ_{ind} – induced potential at the membrane surface. At higher frequencies RC properties must be attributed to the resistances R_{e1}, R_{e2}, and R_i too (Internal and external resistors were split to obtain a reference potential ψ_{ref})

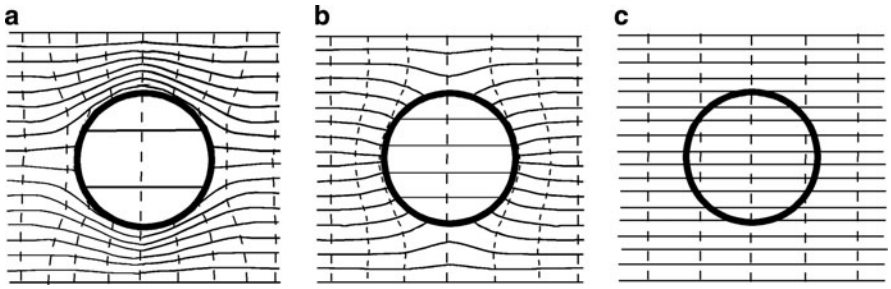

Fig. 3.44 Current lines (——) and equipotential lines (- - -) in, and around a spherical cell in a homogeneous electric AC field. In contrast to Fig. 3.45 the polarization charges at the membrane are not depicted. (**a**) The cell in a low frequency, (**b**) and (**c**) – in a high frequency AC field. In case **B** the conductivity of the external medium is lower than that of the internal one, in case **C** the cell is surrounded by a physiological medium where the permittivities of the internal and external milieus are the same

low-frequency AC fields, because of the high membrane resistance, the field inside the cell is negligible, and the external field becomes deformed (Fig. 3.44a). The membrane capacitor will be increasingly bridged with increasing frequency. This leads to a change of the field distribution in, and around the cell (Fig. 3.44b, c). The field penetration increases with increasing frequencies. In parallel, the membrane polarization decreases, and the polarization of the cytoplasm increases. Taking into account that there is no large difference in the permittivities between medium and cell plasma, the degree of field penetration will be high and constant at frequencies above some 100 MHz.

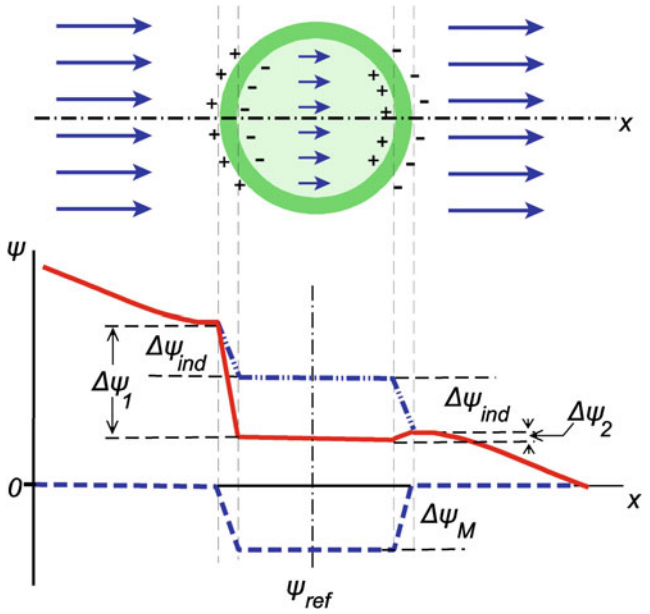

Fig. 3.45 The distribution of charges and potentials at a spherical cell in an external DC field. In the upper part the arrows indicate direction and strengths of the field vectors. The charges near the membrane are the result of polarization by the external field. In the lower part of the figure, the potential profile is depicted over the *x*-axis through the center of the cell: - - - membrane potential without external field influence, including the undisturbed in vivo transmembrane potential ($\Delta\psi_M$); - - - - potential induced by the external field, without the in vivo potential (induced potential difference: $\Delta\psi_{ind}$); —— actual potential function, as the sum of both functions and the resulting membrane potential differences at both poles of the cell ($\Delta\psi_1$, $\Delta\psi_2$). ψ_{ref} is the reference potential according to Fig. 3.43. In contrast to Fig. 2.48, the electric double layer is not taken into account for simplicity

Let us consider the case of Fig. 3.44a in detail in order to discuss the influence of external fields on the membrane potential of a spherical cell in DC or in extremely low-frequency AC fields (Fig. 3.45). Neglecting the conductivity of the membrane, it can be understood that the external field leads to an accumulation of charges, especially in membrane areas that are orientated perpendicular to the undisturbed current lines. This leads to a deformation of the external electric field, and to the charging of the membrane capacitor. The polarization of the membrane induces an intracellular counter-field which significantly mitigates the field influence from outside (short arrows inside the cell). Because of the low field strength inside the cell, a possible polarization of the membranes of cell organelles can be neglected, at least at DC, or low-frequency AC fields.

The transmembrane potential difference $\Delta\psi_{ind}$ induced by an external field, corresponds to the difference $\psi_{ref} - \psi_{ind}$ in Fig. 3.43. There is no potential difference across the resistor R_i and no current flows inside the cell.

Because of the deformation of the external field by cell polarization (Fig. 3.44a), the potential ψ_{ind} (Fig. 3.43) is somewhat different from the potential at an identical x-coordinate away from the cell. This was taken into consideration introducing the resistor R_{el} and a bending of the extracellular potential function near the membranes in the lower part of Fig. 3.45. The extent of this deviation depends on the shape and the radius (r) of the cell. In the case of a spherical cell, the characteristic distance from the center of the spherical cell is $1.5r$.

The polarity of the induced potential difference ($\Delta\psi_{ind}$) at both sides of the cell corresponds to the external field, whereas the undisturbed in vivo transmembrane potential $\Delta\psi_M$ is always oriented inside-out. Therefore, the cell is polarized in the same direction as the induced potential difference ($\Delta\psi_{ind}$) on one side, and oriented in an opposite direction on the other side. This means that the resulting potential differences ($\Delta\psi_1$, $\Delta\psi_2$) differ from one another on both sides of the cell. The in vivo potential at locations oriented perpendicular to the field lines will not be influenced by the external field at all. For this reason, the reference level for the induced potential (ψ_{ref}) is identical inside and outside the cell in Fig. 3.45.

To calculate the induced membrane potential (ψ_{ind}), a small column can be considered. It is oriented in the field direction and cut out along a line through the center of the sphere with the cross-sectional area A. The characteristic length up to which cell polarization may enhance the external medium potential is $r_e = 1.5r$. Furthermore, we shall consider that the Ohmic current flow through the membrane resistor R_m (Fig. 3.43) can be neglected. In this case the following relations apply:

$$R_i = \frac{r}{g_i A}; \quad R_{e1} = \frac{0.5r}{g_e A} = \frac{r}{2g_e A}; \quad C_m = \frac{C}{A} \tag{3.213}$$

Furthermore, the time constant (τ) can be considered reflecting the characteristic time to charge the membrane capacitor:

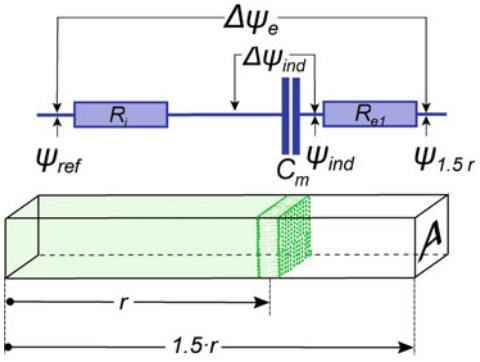

Fig. 3.46 A column with a constant cross-sectional area A, cut out from the cell (Fig. 3.43), and the electric scheme for the corresponding circuit. The resistor R_m and the capacitors for the cytoplasmic and external media are neglected (Corresponding to the approach of Gimsa and Wachner 1999)

$$\tau = C(R_i + R_{e1}) = r\ C_m\left(\frac{1}{g_i} + \frac{1}{2g_2}\right) \tag{3.214}$$

Considering an external field of strength **E**, the potential difference outside the cell, from point ψ_{ref} to $\psi_{1.5\ r}$ is: $1.5r\mathbf{E}$. The same potential must drop over the column. Using the proportionality between the impedance, as effective AC resistance, and the potential drop, and considering the impedance of the membrane capacitor: $Z^* = 1/Y^* = -j/\omega C$, we get:

$$\frac{\Delta\psi_{ind}}{1.5\mathrm{E}r} = \frac{-j/\omega C}{-j/\omega C + R_1 + R_{e1}} = \frac{-j}{-j + \omega C(R_1 + R_{e1})} \tag{3.215}$$

Introducing the time constant of Eq. 3.214 and rearranging, leads to:

$$\frac{\Delta\psi_{\mathrm{ind}}}{1.5\mathrm{E}r} = \frac{-j}{-j + \omega\tau} \tag{3.216}$$

Noticing that:

$$|-j + a| = \sqrt{1 + a^2}$$

We obtain for the absolute value of the induced transmembrane potential:

$$\frac{\Delta\psi'_{\mathrm{ind}}}{1.5\mathrm{E}r} = \frac{1}{\sqrt{1 + \omega^2\tau^2}} \tag{3.217}$$

After rearrangement and introduction of the expression τ from Eq. 3.214 one gets:

$$\Delta\psi_{\mathrm{ind}} = \frac{1.5\,\mathrm{r}\,\mathbf{E}}{\sqrt{1 + \left[\mathrm{r}\omega\mathrm{C_m}\left(\frac{1}{\mathrm{g_i}} + \frac{1}{2\mathrm{g_e}}\right)\right]^2}} \tag{3.218}$$

This is the maximum of the membrane potential induced in the direction of the external field. At an angle perpendicular to the field (Fig. 3.45) the induced transmembrane potential vanishes. To consider the induced potential at all points of the sphere, the radial coordinate α can be introduced, and Eq. 3.218 must be multiplied by $\cos\alpha$.

Using common cell parameters, like for example: $r = 10^{-5}$ m, $C_{sp} = 10^{-2}$ F m^{-2}, $g_i = 0.5$ S m^{-1}, and $g_e = 1$ Sm^{-1} in Eq. 3.217 a relaxation time of $\tau = 2.5{\cdot}10^{-7}$ s is obtained. Introducing this parameter into Eq. 3.18 it is easy to demonstrate that for low-frequency AC fields ($\nu < 10^5$ Hz), the denominator of this equation will approach 1.

For DC fields and low-frequency AC fields, including the dependence of the vector angle α, the equation reduces to:

$$\Delta\psi_{\text{ind}} = 1,5\ \text{E}r\cos\alpha \tag{3.219}$$

As already mentioned, this, as well as Eq. 3.218, is correct only for a very high membrane resistance which is justified in most cases. Using this equation one can for example calculate that in a spherical cell with a diameter $r = 10$ µm a superimposed transmembrane potential of $\Delta\psi_{\text{rel}} = 1.5$ V will be induced at position $\alpha = 0°$ by an external low frequency, or DC field of approximately $\mathbf{E} = 100\ \text{kV m}^{-1}$.

As will be considered in the next section, this equation is useful to calculate the field strength which is necessary to manipulate cells by electric membrane break down or cell-cell fusion. Conversely, it must be noted that it is not applicable to the calculation of stimulus thresholds for muscle and nerve tissues (see Sect. 4.6.2). In this case, there are complicated situations of field distribution in the intercellular space. Furthermore, these cells are extremely elongated. Finally, in some cases they are electrically connected to one another by gap junctions. In this case, not the length of the individual cell in the field direction is representative of the induced membrane potential, but the length of the whole electrically connected system.

Further Reading

Gimsa and Wachner 1999; Grosse and Schwan 1992.

3.5.5 Manipulation of Cells by Electric Fields

The interaction of cells with external electric fields as discussed in the previous section has led to various applications in biotechnology. In contrast to the effects of weak electric and electromagnetic fields on biological systems, which we will discuss later (Sects. 4.6, 4.7), these applications require rather strong fields. Field-induced charges that accumulate at cell–medium interfaces on the one hand influence the transmembrane potential ($\Delta\psi$), and on the other directly induce mechanical forces which are able to move or deform the cells.

The application of strong electric fields in physiological solutions with considerable conductivities will, of course, induce significant Joule-heating. The electrical power, absorbed in a medium per volume equals $\mathbf{E}^2 g$ (see Eq. 4.23). For a specific conductivity in a physiological milieu of approximately $g = 1\ \text{S m}^{-1}$, the applied field strength of $\mathbf{E} = 10^5\ \text{V m}^{-1}$, as required for these applications, results in an absorbed power density of $10^{10}\ \text{W m}^{-3}$, i.e., $10^7\ \text{W kg}^{-1}$. This leads to an enormous local heating of the system which can only be avoided by the use of short field pulses, an artificial medium of low conductivity, or microscopic electrode systems with relatively large heat-conducting surfaces.

Let us first consider the influence of strong electric fields on the hyper-, or hypopolarization of the membrane. As explained in the previous section, an externally applied DC field increases the membrane potential ($\Delta\psi_1$ in Fig. 3.45) on one side of the cell, and consequently, raises the internal electric field strength in the membrane at this point. As mentioned in Sect. 2.2.1, the electric transmembrane field in vivo is nearly 10^7 V m^{-1}. If the transmembrane potential is artificially elevated, reaching an amount of approximately one volt, the internal field strength becomes overcritical, and the membrane will be destabilized. In this so-called *electric break down* the membrane loses its property as a diffusion barrier. Its electrical resistance breaks down, which may even lead to complete destruction of the cell by subsequent lysis. However, by selecting the parameters of treatment properly, i.e., using moderate shape, amplitude, and duration of the pulse at optimal temperature and electrolyte conditions, a reversible electric break down can be induced. This additionally requires use of the proper composition of the external medium to avoid Donnan-osmotic cytolysis of the cells (see Sect. 3.2.5). In this case, the membrane may stabilize again and the cell may survive. The duration of the applied field pulses is of the order of 1–500 μs.

The electric break down of cell membranes is used for so-called *electroinjection* or *electropermeabilization*. When cells are suspended in a solution containing macromolecules or small particles, the short moment of the induced membrane destabilization may be sufficient to allow the penetration of these molecules or particles into the cell with sufficient probability. Additionally, such disturbances of the membrane structure can also increase the ability of the cell to absorb particles by phagocytosis. Electroinjection is used for example to introduce drugs, genetic materials, or even small organelles into cells.

The basic equation for the external field which is necessary to reach the required membrane field for these manipulations was introduced in the previous section (Eqs. 3.218, 3.219). It indicates that the external field strength ($\mathbf{E}$) which is required to induce an electric break down is inversely proportional to the cell radius (r). Stronger external fields therefore are needed for smaller cells. Recently, conditions were investigated to achieve break down also in intercellular organelles, such as mitochondria or chloroplasts. For this, nanosecond pulses are used with field strengths up to 10^7 V m^{-1}.

If two cells are attached to each other during application of an electric pulse, or a short time after the pulse, the points of induced membrane instability may interact with each other, leading to a fusion of the lipid phases of the membranes of the neighboring cells. Subsequently membrane fusion, an osmotically governed process of fusion of the whole cell may follow. This process of *electrofusion* of cells has recently been explained by several mechanisms which cannot be discussed here in detail. Membrane and subsequently cell fusion are in fact triggered by electric pulses, but need a time longer than the pulse duration to be accomplished.

Electrofusion of cells is broadly applied in biotechnology. In contrast to the induction of cell fusion by some chemical agents, electrofusion has the advantage of being much more selective. Using sufficiently small electrodes, it is possible to fuse even two particular cells. Meanwhile, methods have been developed to bring

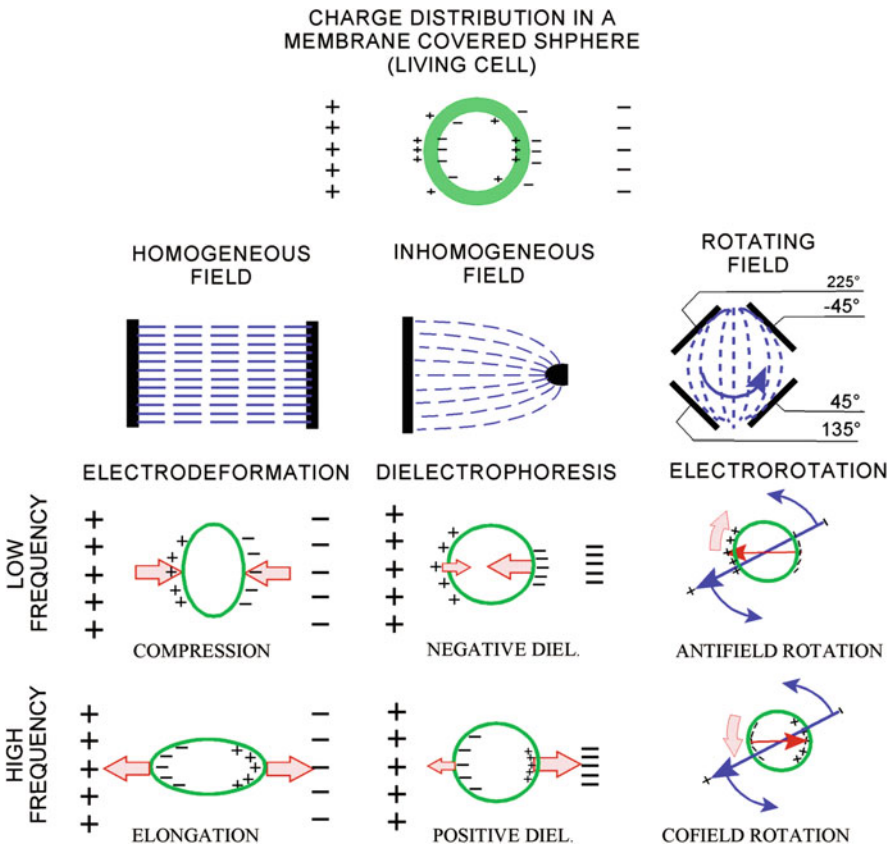

Fig. 3.47 Various electromechanical effects on a spherical cell model. Above: general distributions of charges near the membrane in relation to electrode charges. Below: snapshots of charge distributions in homogeneous, inhomogeneous, and rotating AC-fields at different frequencies

specific cells into contact with each other. Usually, an aggregation by dielectrophoresis is used for this purpose.

Charge separation, i.e., the induction of cell dipoles of various quality, leads to a number of electromechanical phenomena. These can be used either to separate and mechanically manipulate cells, or to directly measure dielectric properties of individual cells (Fig. 3.47).

In general, two qualitative cases of cell polarization exist. At frequencies below membrane dispersion the effective charge distribution is opposite to the case at higher frequencies, where the membrane is fully capacitively bridged. Corresponding to this, at low frequencies cells will be compressed by the field, whereas they will be stretched in high frequency fields.

In inhomogeneous fields, the cell hemisphere in the higher field area experiences a higher polarization, i.e., a higher force than its opposite. This imbalance leads to a

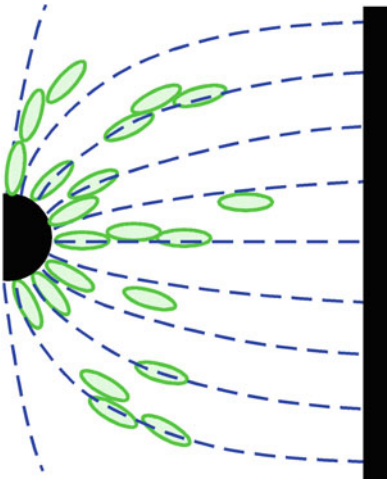

Fig. 3.48 The dielectrophoresis of yeast cells in an inhomogeneous field which spans between the peak of a needle, shown as a small hemispheric electrode, and a flat electrode opposite. The figure shows the accumulation of the cells near this electrode as a result of positive dielectrophoresis. The polarization of the cells additionally leads to their mutual attraction and the *pearl chain formation*

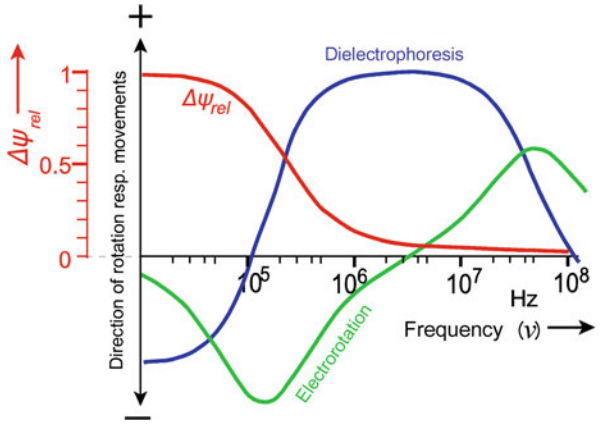

Fig. 3.49 An example of electrorotation (green line) and dielectrophoresis (blue line) of a cell in a solution of low conductivity, dependent on the applied frequency. Positive values of the electrorotation spectrum mean spinning in the direction of field rotation (cofield rotation), or in the case of dielectrophoresis a movement into the direction of higher field intensity, and vice versa. Additionally, the parameter $\Delta\psi_{\mathrm{rel}} = [1 + (2\pi\nu)^2]^{-1/2}$ is depicted as a function of frequency (red line) according to Eq. 3.218. It describes the influence of the induced membrane potential (with the help of Wachner and Gimsa)

translocation of the cell which is called *dielectrophoresis* (Fig. 3.48). The direction of the dielectrophoretic translation depends on the orientation of the effective polarization of the cell and on the orientation of the field gradient. According to the polarization effects there are frequency regions of *negative* as well as of *positive* dielectrophoresis (see Fig. 3.49). The driving forces in all electromechanical processes are proportional to the square of the field strength ($\mathbf{E}^2$). Because of the deformation of the field around the individual cells (see Fig. 3.44), the gradient of

the fields directly near the cells results in mutual dielectrophoresis, i.e., in the attraction of cells to each other. This leads to the formation of pearl chain-like structures. This effect can of course occur also in homogeneous external fields.

As depicted in Fig. 3.47 in rotating fields the dipole induction results in *electrorotation*. Rotating fields can be applied using multielectrode systems. In the case of a four-electrode system, as shown in this scheme, a generator produces an AC signal with adjustable frequency, which is split into four signals of particular phase relations. In this case, the field rotates at the frequency of the generator. Correspondingly, the induced dipole rotates too. Because of the time constant of polarization, an angular difference however may occur in between the induced dipole, and the external field vector. This leads to a permanent force of interaction between them, either in the form of repulsion, resulting in an *antifield rotation* at low frequencies, or attraction at higher frequencies, the so-called *cofield rotation*. The resulting rotation frequency of the cells in the chamber is much smaller than that of the field, and again is proportional to the square of the applied field strength. In contrast to dielectrophoresis, which can be described by the real part of the induced dipole, electrorotation is related to its imaginary component.

In Fig. 3.49, as an example, the frequency dependence of electrorotation of a cell in a low conductivity medium is depicted. It shows that the antifield rotation occurs at frequencies of about 100 kHz, followed by a point of zero rotation and finally, in the MHz region a maximum of cofield rotation occurs at about 50 MHz.

Figure 3.47 represents the simplest case, a single shell model, which considers a dielectrically homogeneous sphere, covered by a membrane. It represents a spherical cell without electrically significant organelles (Fig. 3.43). Non-nucleated swollen erythrocytes can be described by this model. In this case, the radius of the cell and the membrane thickness are included as geometrical parameters in the corresponding equations, as well as the conductivities and permittivities of the membrane, the cytoplasm, and the external medium. These model considerations indicate that the peak of the antifield rotation in its frequency position, and its amplitude, reflects the membrane properties of the cell. This first characteristic frequency is determined by the time constant given by Eq. 3.214. Changing the permeability of the membrane as a diffusion barrier, this peak vanishes. The maximum of the cofield rotation, i.e., the second characteristic frequency, indicates the conductivity and permittivity of the cytoplasm in relation to that of the external milieu.

Electrorotation and dielectrophoresis are approved methods to measure dielectric properties of individual living cells. There are a number of investigations indicating that electrorotation can measure alterations of cell properties which are induced by drugs, toxic agents, virus attacks, cell activations, and other events. Under special conditions even properties of organelles like nuclei or vacuoles can be measured. Automatic video systems and methods of dynamic light scattering are applied to register the induced movement.

In contrast to electrorotation, which is mostly used for cell analyses, dielectrophoresis can also be applied for preparative cell separation and other

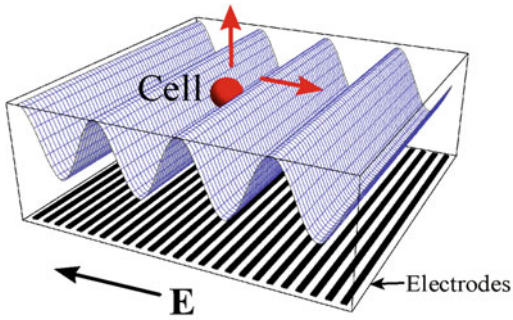

Fig. 3.50 Cells moving in a system of interdigitated travelling-wave electrodes. Note that the direction of field propagation is opposite to the direction of cell motion according to negative dielectrophoresis (After Fuhr et al. 1996)

biotechnologically interesting techniques. As already mentioned, the formation of cell chains, or cell-cell attachment by dielectrophoresis is used to perform selective electrofusion of cells by applying short pulses of high field intensity. Recently, special microdevices have been constructed to investigate single cells by electrorotation and to manipulate them by dielectrophoretic translations. According to the large relative surface of such chambers in relation to their volume, heating of the samples was minimized. This allows one to apply dielectrophoresis and electrorotation also in physiological solutions of relatively high conductivities. It is also possible to produce microscopic field traps, in which cells, lifted by dielectrophoretic force can be held in a stable position without any surface contact. Using electrode arrangements inducing traveling waves (Fig. 3.50), cells can be moved on microchips. This new technique opens enormous possibilities for biotechnological applications.

Further Reading

For electrorotation and dielectrophoresis: Georgiewa et al. 1998, Gimsa and Wachner 1998; Fuhr et al. 1996; Fuhr and Hagedorn 1996; electromanipulation and electrofusion: Cevc 1990; Kolb et al. 2006; Lynch and Davey 1996, Zimmermann 1996; dielectric cell deformation: Sukhorukov et al. 1998; Ziemann et al. 1994.

3.6 Mechanical Properties of Biological Materials

Biomechanics is a branch of biophysics that examines mechanical properties of biological materials and forces acting upon and within the biological structure, as well as effects produced by such forces. It explains the anatomical stability of plants and animals, the movements of limbs, and in this way the mechanics of walking, flying, and swimming. A special branch of biomechanics, biorheology concerns the mechanics of blood flow and other kinds of fluid movement inside the organism. Hearing and other kinds of acoustic organs can be explained on the basis of various kinds of cellular mechanoreceptors.

Biomechanics is in fact one of the oldest branches of biophysics. Its development followed closely that of physical mechanics itself. In the period of renaissance the pioneers of mechanics, such as Gallileo Galilei, René Descartes, Isaac Newton, and many others were always also interested in the mechanics of animals. The first classical book on biomechanics, Alfonso Borelli's "De motu animalium" was printed in Rome in 1680. It already contained basic physical considerations on swimming, flying, and movement of animals as well as various calculations of moments of human limbs and the spine under conditions of loads. As we already mentioned in Sect. 1, this book marked the beginning of medical physics, which was called at that time "iatrophysics." Furthermore, D'Arcy Thompson's book "On Growth and Form," published first in 1917 must be mentioned. This book analyzed for the first time basic processes of cell mechanics, shape formations, and many other biomechanical fields.

In recent decades biomechanics has become increasingly important in diagnostics and therapeutics, especially the biophysics of blood circulation (hemorheology), including the properties of blood vessels and the pumping properties of the heart, the biomechanics of the skeleton, as well as of limbs and joints, all of which form the basis for medical physics and medical engineering.

A further interest in special questions of biomechanics comes from sport. Here various kinds of complex body motions are of interest to optimize outcome. There is also interest in biomechanics from ecology, considering life in moving fluids, wind resistance of plants, etc.

Further Reading

On biomechanics in general, Alexander 2003; Bels et al. 2003; Chein et al. 2008; Fung 1993; Nigg and Herzog 2007; Niklas 1992; Oomens et al. 2009; Özkaya and Nordin 1999; Skalak and Chien 1987; Vogel 1994.

3.6.1 Some Basic Properties of Fluids

Many parameters characterizing the physical property of a fluid are defined for the case of laminar flow. A flow is called *laminar* if particles of the fluid move in parallel layers to each other (for more on the properties of laminar flow see Sect. 3.7.1). Figure 3.51 illustrates examples of three kinds of laminar flow. In contrast to turbulent flow, processes in laminar flows can be calculated by linear thermodynamic approaches (see Sect. 3.1.3, Fig. 3.3).

To understand the approaches of fluid mechanics, and to characterize laminar fluid profiles, we first need to introduce the *velocity gradient* (γ), also known as *shear rate*. It is the derivation of the streaming velocity ($\mathbf{v}$) with respect to a coordinate (z), perpendicular to the direction of the flow.

$$\gamma = \frac{\mathrm{d}\mathbf{V}}{\mathrm{d}z} \tag{3.220}$$

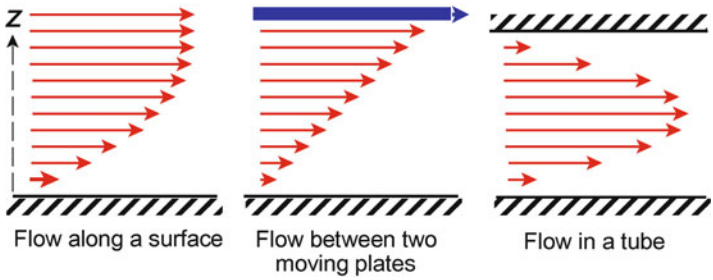

Fig. 3.51 Various examples of laminar velocity profiles

The measure of this velocity gradient is therefore s^{-1}.

The shear rate of a moving fluid is at a maximum near a solid surface, or in the case of a tube, near the wall. Far away from the surface this gradient becomes zero.

If two parallel plates slowly move one in relation to the other, they generate a laminar flow in between, the velocity gradient of which is always constant, and is equal to the relative velocity of the plates, divided by their mutual distance ($\gamma = \Delta\mathbf{v}/\Delta z$).

The force (**F**) driving a plate with a surface A in the case of laminar flow is proportional to the velocity gradient (γ) between the two plates:

$$\mathbf{F} = \eta\gamma\, A \tag{3.221}$$

In this equation a friction coefficient (η) is introduced which is called *viscosity*. The viscosity therefore determines the force, which is required to move a plate with an area of 1 m^2, at 1 m distance from a parallel surface at a velocity of 1 ms^{-1}, if a laminar flow of a liquid between these surfaces is induced. Thus, the measuring unit for the viscosity is: Nsm^{-2}, or: Pa · s. Sometimes an older unit P (Poise) is used, whereas: 1 P = 0.1 Nsm^{-2}.

Parallel to the viscosity (η), another parameter, the *fluidity* ($\varphi = 1/\eta$) is used, as well as the *kinematic viscosity* ($\nu = \eta/\rho$), a parameter, containing the density (ρ) of the fluid. The force deforming a body in a streaming fluid with a velocity gradient (γ) is given by the *sheer stress* (τ):

$$\tau = \eta\gamma \tag{3.222}$$

The viscosity depends to a high degree on the temperature. Especially for aqueous solutions this is caused by the cluster structure of the water (see Sect. 2.2.2). In contrast to the viscosity of pure water at $T = 0°C$, which is 1.79 mPa s, it amounts at 25°C to only 0.89 mPa s, and at 100°C, finally to 0.28 mPa s (see Fig. 2.16 in Sect. 2.2.2).

In order to characterize the influence of dissolved or suspended substances on the viscosity of a fluid, the following derived parameters are used:

Relative viscosity: $\eta_{\text{rel}} = \frac{\eta}{\eta_W}$
Specific viscosity: $\eta_{\text{sp}} = \eta_{\text{rel}} - 1$
Reduced viscosity: $\eta_{\text{red}} = \frac{\eta_{\text{sp}}}{c}$ (in: 1 mol^{-1})
Intrinsic viscosity: $[\eta] = \lim_{c \to 0} \eta_{\text{red}}$ (in: 1 mol^{-1})

where η is the viscosity of the solution of suspension, η_W is the viscosity of the pure solvent, and c is the molar concentration of the solute.

The viscosity increases with increasing concentration of the dissolved or suspended substances. As already pointed out in the definition of the intrinsic viscosity $[\eta]$, even the reduced viscosity of a solution is a function of the concentration. The intrinsic viscosity contains information on the structure and the molecular mass of a substance.

For diluted suspensions of small rigid spherical particles the *Einstein relation* can be applied:

$$\eta_{\text{sp}} = 2\,5\,V_{\text{rel}} \quad (\text{for}: \quad V_{\text{rel}} < 0.1) \tag{3.223}$$

The relative volume (V_{rel}) is the volume of all particles in the suspension together in relation to the volume of the suspension. For suspension of cells (sperms, erythrocytes, etc.) the term *cytocrit* (or specifically *spermatocrit, hematocrit*) is used. It should be emphasized that neither the absolute size of an individual particle, nor the homogeneity of the diameters of all particles in the suspension are of importance for this relation. The Einstein equation, however, is correct only for very diluted suspensions.

Fluids, the viscosity of which is independent of the velocity gradient (γ) are called *Newtonian fluids*. In contrast, *non-Newtonian fluids* alter their viscosity depending on this parameter.

In Fig. 3.52 the behavior of various kinds of non-Newtonian fluids are demonstrated. *Dilatant* fluids are mostly suspensions of solids, like quartz particles. They produce entropy by mutual friction as much as the shear rate of fluid increases. The *Bingham-plastic* behavior occurs, for example, in a suspension of

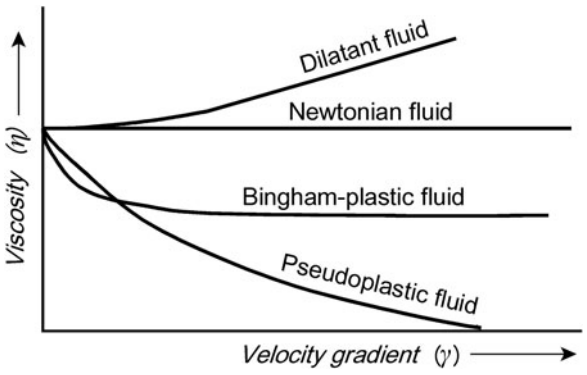

Fig. 3.52 The dependence of the viscosity (η) on the velocity gradient (γ) for a Newtonian fluid (——), and for various types of non-Newtonian fluids (——) (After Glaser 1989)

nonspherical particles. In this case velocity gradients lead to their orientation, which decreases the viscosity of the suspension. At certain points, if the particles are oriented at maximum these suspensions behave like Newtonian fluids. The same behavior is to be expected if the particles tend to aggregate in the resting fluid, but disaggregate at low shear stress.

The most common property of biological fluids is the *pseudoplastic* behavior. It occurs for example in blood (Sect. 3.6.2, Fig. 3.54), and many other biological fluids with heterogeneous composition. Different components of these fluids, such as for example blood cells, proteins and other macromolecules, aggregate, orientate, and deform at various shear gradients. The resulting function, therefore, does not come to a saturation level at reasonable values of γ.

These shear-induced processes of course need some time to become established. In the case of spontaneous induction of a shear gradient, or of a sudden change of it, a time dependence of the viscosity occurs. This behavior was first observed in 1923 in gels, which could be transformed by shaking into a liquid sol. The same property was also found in the viscous behavior of cell protoplasm. The term "thixotropy" was introduced as a combination of the Greek words thixis (stirring, shaking) and trepo (turning or changing). According to the IUPAC terminology: *thixotropy* is defined as the continuous decrease of viscosity with time when flow is applied to a sample that has been previously at rest and the subsequent recovery of viscosity in time when the flow is discontinued. Confusion between thixotropy and shear thinning still persists in some cases in the literature. Therefore, it should be emphasized that thixotropy applies to the time dependence of non-Newtonian behavior of a liquid.

The opposite phenomenon also exists, i.e., a reversible, time-dependent increase in viscosity. This could be the result of flow-induced aggregation of particles. This is called *antithixotropy*, earlier known as *rheopexy*. In the next section we will demonstrate these properties for the case of the behavior of blood.

These properties of fluids must be taken into account, when choosing instruments to measure the viscosity. Some of the most common methods are depicted in Fig. 3.53. In the case of a capillary viscosimeter (Fig. 3.53a) the time is measured which a given fluid needs to pass a capillary under a known pressure. For this, a certain volume of the fluid is placed in the left part of the U-tube and the time is measured for the fluid to pass the two marks M_1 and M_2. This time is proportional to the viscosity. In this way the viscosity can be measured after calibration of the setup using a fluid with known viscosity. This so-called Ostwald viscosimeter was later modified by Ubbelohde in such a way that a further vertical tube arranged at the end of the capillary leads to an interruption of the fluid. Another way, to measure the viscosity of a fluid is the use of falling bodies (Fig. 3.53b). Both methods have the advantage of being simple and in relation to other equipment, inexpensive. However, conversely they are only suited to measuring the viscosity of Newtonian fluids because the velocity gradient of the streaming fluid in the capillary, as well as between the falling sphere and the wall of the tubes, are not at all constant (see Fig. 3.51).

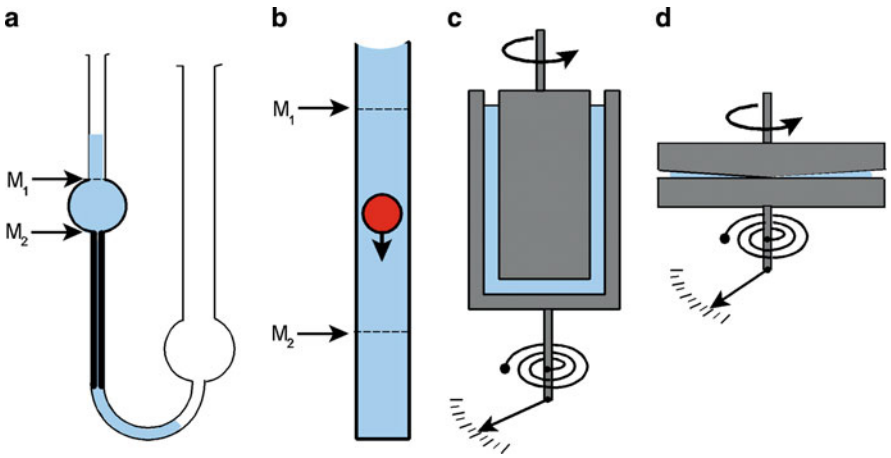

Fig. 3.53 Various setups to measure the viscosity of fluids. (**a**) Capillary viscosimeter after Ostwald, (**b**) viscosimeter with falling sphere, (**c**) coaxial-type rotational viscosimeter, (**d**) cone-type rotational viscosimeter

To investigate the viscosity of non-Newtonian fluids measuring instruments are required which allow one to establish well-defined shear gradients. As demonstrated in Fig. 3.51, this is possible between plates moving parallel to each other. This principle is used in so-called *Couette*, or *rotational viscosimeters*. In this case the fluid is placed in the gap between two coaxial cylinders (Fig. 3.53c), or between a flat plate and a truncated cone (Fig. 3.53d). By moving one part of this equipment, the viscosity of the fluid transmits a torque which can be measured. Usually, one part is rotated with adjustable speed, and the torque of the opposite part is measured by a sensitive instrument. By changing the speed and the thickness of the cleft, one can produce various velocity gradients. In the case of the rotating cone the tangential velocity is compensated by the increasing broadness of the cleft. Therefore, in this case a constant velocity gradient is also established. Curves like those shown in Figs. 3.52 and 3.54 are produced using such instruments.

Recently, the *Stabinger viscosimeter* has been used as a modification of the classic Couette rotational viscosimeter. In this case, the internal cylinder is hollow and thus floats freely in the sample, centered by centrifugal forces. This avoids any bearing friction. The torque of this cylinder is implemented by a rotating magnetic field.

Further Reading

On thixotropy, Mewis and Wagner (2009).

3.6.2 The Viscosity of Biological Fluids

Through the development of sensible rotational viscosimeters it was possible to investigate viscous properties of a large number of biological materials, such as

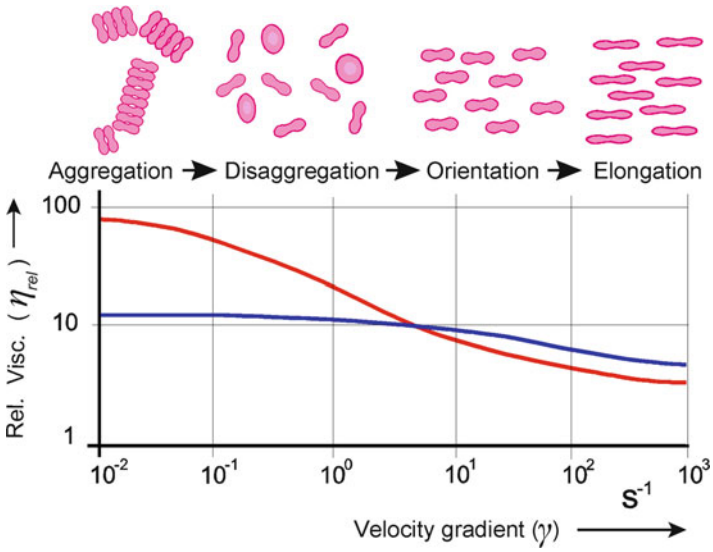

Fig. 3.54 The relative viscosity (η_{rel}) dependent on the velocity gradient (γ) of human blood (——), and heat-hardened erythrocytes, which were resuspended in plasma (——). The difference of both curves in the region of low shear rates is achieved by aggregation, disaggregation and elongation of the native red blood cells at increased shear rates. (Data from Lerche and Bäumler 1984)

blood, lymph, various secretions, synovial fluids, and many others. This has led on one hand to a better understanding of the processes of blood flow, of the mechanics of joints, etc., and on the other hand it has become a useful indicator in diagnostics.

In Fig. 3.54 the viscosity of a suspension of red blood cells is depicted as a function of the velocity gradient. In contrast to blood plasma which appears to be a Newtonian fluid, these suspensions, as well as the whole blood, indicate pseudoplastic thixotropic behavior. The reason for this is complex: Native erythrocytes, suspended in blood plasma aggregate at low shear rates. These aggregates are not very stable and disaggregate at somewhat higher shear rates. This disaggregation lowers the viscosity of the suspension. A further increase of the shear rate leads to orientation, and finally to a deformation of the erythrocytes, further decreasing the viscosity. Looking at the behavior of erythrocytes which were hardened by fixation, neither the aggregation, nor deformation caused by shear stress occurs. Living erythrocytes in the shear gradient become elongated ellipsoids that are oriented along the streaming vectors. With increasing shear stress they become more and more elongated. Whereas slow elongations are reversible up to a certain degree of elongation, irreversible alterations in the membrane occur. Eventually hemolysis occurs as a result of maximal shear stress.

The energy of the streaming fluid which leads to deformation of the cells corresponds to the shear stress (τ) according to Eq. 3.222. This means that it depends not only on the velocity gradient (γ), but additionally on the viscosity (η)

of the fluid. To investigate shear-induced shape deformations therefore, high viscosity solutions are usually used, for example, solutions of high molecular weight dextrans of various concentrations.

For biomechanical problems in orthopedics the properties of the synovial fluid which is located in the joint cavity and surrounded by the joint capsule are of particular interest. In the language of tribology, which represents the effects of friction in moving machines, the mechanism of joints represents a kind of *depot lubrication* with porous surfaces. The expression "depot lubrication" points to the synovial fluid which is accumulated in the bursa of the joints. "Porous surface" refers to the articular cartilage which covers the cortical bone in the joint capsule by a 0.3–0.5-mm thick layer. The joints are not only burdened by movement, i.e., by a shear stress of the synovial fluid, but additionally in some cases by a considerable static pressure. It must be guaranteed that these loads do not press the synovial fluid out of the cleft. In fact the synovial fluid has thixotropic pseudoplastic properties. A large viscosity of this fluid in the joints (η between 1 and 40 Pa s) prevents its exclusion from the cleft by hydrostatic pressure. If the joint is moving however, shear rates up to 10^5 s^{-1} appear. In this case, the viscosity of the synovial fluid decreases to 10^{-2} Pa s, leading to a highly effective lubrication of the joint. This particular property of synovial fluid is caused by a special structure of proteoglycans, which are high molecular weight glycoproteins with an intriguing structure.

This leads us to considerations of the viscosity of microscopic, or even supramolecular structures. It must be pointed out that the definition of the viscosity, as given in the previous Sect. 3.6.1, again comes from continuum physics. It does not take into account the behavior of molecules and supramolecular composition of the liquids. It just considers a homogeneous, continuous fluid without any structures. Already the non-Newtonian behavior discussed above indicates some limitations of this assumption. Moreover, problems arise, if we consider viscoelastic properties of cells and their constituents. At the beginning of the twentieth century, at a time when the physical properties of cells were first being discussed, many investigators measured the viscosity of the cytoplasm in relation to various biological functions. This question from a modern point of view is outdated because of our knowledge of the highly organized structure of this region. Even in cells with pronounced cytoplasmic streaming as in amoebas or in plant cells the question of the origin of these movements and of the molecular mechanisms driving these flows, is of more central interest than phenomenological models using plasma viscosity in general.

The problem of viscosity in molecular and supramolecular dimensions is important to a high degree in relation to the "viscosity" of the cell membrane (see also Sect. 2.3.4). Various molecular probes allow us to obtain physical parameters of the membrane, which in any case are functions of the viscosity. So, for example, a lateral diffusion constant of fluorescent labels in the membrane can be measured. Using the Einstein equation (Eq. 2.39) some conclusions can be made on the viscosity of their surrounding medium. In the same way the rotation diffusion constant of specific spin-probes bound to particular membrane molecules, measured by electron-spin resonance techniques (ESR), allows us to determine the viscosity. Furthermore, microscopic stress deformations can be applied.

In all cases the viscosity which is determined by these methods is a typical effective parameter. In the same way as for example the hydration radius of an ion (see Sect. 2.2.2), this parameter depends on the techniques and the physical phenomenon used for its determination. One should not consider these "quasi-viscosities" in the same sense as the phenomenological viscosity of a fluid which we discussed before. Moreover, the mechanical anisotropy of the membrane must be taken into account, which means that the mobility of particles is not equal in all directions of space.

Further Reading

Leyton 1975; Owens 2006; Skalak and Chien 1987.

3.6.3 Viscoelastic Properties of Biomaterials

The simplest kind of deformation of a body is its stretching, hence the basic parameters of viscoelasticity are defined for this case. In Fig. 3.55 this is illustrated by a stress–strain diagram, the coordinates of which are defined as follows:

$$\text{Stress:} \quad \sigma = \frac{\mathbf{F}}{A} \quad (\text{in}: \ \text{N m}^{-2}) \tag{3.224}$$

$$\text{Strain}: \quad \varepsilon = \frac{\Delta l}{l} \tag{3.225}$$

where: **F** is the force, A is the cross-sectional area of the body, l its length, and Δl is the difference between the resting and the extended material. The strain (ε) is just a relation and therefore does not have a measuring unit.

In general this diagram indicates some regions of different behavior: In the region of minimal strain up to the limit of proportionality (P) *Hooke's law* is valid stating that the strain (ε) is directly proportional to the stress (σ). The ratio of these two properties is called the *modulus of elasticity*, or *Young's modulus* (Y).

$$Y = \frac{\sigma}{\varepsilon} \tag{3.226}$$

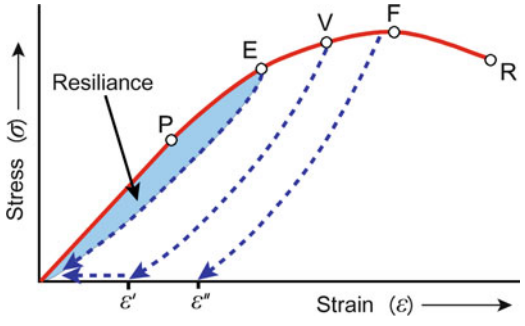

Fig. 3.55 A generalized stress–strain diagram. P – limit of proportionality, E – limit of elasticity, V – limit of reversible viscoelastic deformation, F – point of floating deformation, R – rupture point, ε′, ε″ – residual strains, *blue area* – resiliance

Young's modulus (Y) has a measuring unit N m^{-2} or Pa. Because of the large amount of this parameter in common materials, mostly the units 1 GPa $= 10^3$ MPa $= 10^9$ Pa are used.

The linear relationship between stress and strain does not hold for large stress behind point P. The deformation however, is reversible up to the elastic limit (E). This means that the body will spontaneously and quickly return to its original length when the deforming force is removed. In this case, however, the relaxation curve (blue dotted line) did not follow the extension curve (red line). The area between these two lines is called *resiliance*. In general the area in this plot, i.e., the product of stress and strain has the unit of energy. The resiliance therefore represents the thermal energy which is dissipated during the process of extension and relaxation. In terms of irreversible thermodynamics it is the dissipation function ($\mathbf{\Phi}$) of this process (see: Eq. 3.64, Sect. 3.1.4). It results from the viscous friction within the body and is an expression of the nonideal behavior of the system.

If the strain is taken beyond the elastic limit (E), the body begins to deform irreversibly. Up to the limit of viscoelastic deformation (point V), the body quickly relaxes to some residual strain (ε') which eventually may slowly vanish. Overcoming this point, an irreversible deformation (ε'') persists after the stress has been removed. A further extension leads to the point where the body begins to show spontaneous flowing elongation, in spite of further increase in stress until it rips up at point R.

This stress–strain diagram indicates that the deformation of a body outside the limit of proportionality depends not only on elastic, but also on viscose properties of the material. For this the term *viscoelasticity* is used. Biological tissue shows viscoelasticity to a great extent. In this case not only the stationary stress–strain function is important, as demonstrated in Fig. 3.55, but also the kinetics of deformation. It is possible to simulate the strain behavior of elastic and viscoelastic materials by mechanical systems made up of elasticity elements, as well as of viscosity, or damping elements (see Fig. 3.56). In contrast to the spring, which, if it is massless, elongates immediately if a rectangular force function is applied, the viscous damping element (e.g., a dashpot) elongates with a constant velocity and remains in position if the force vanishes. A damping element such as a dashpot combined in series with a spring is called a *Maxwell element*. In the case of an applied rectangular function of force, the spring elongates immediately, then a further continuous elongation of the system proceeds. If the force vanishes, the spring contracts but the viscous part of the Maxwell element remains elongated. If a spring and a damping element such as a dashpot are connected in a parallel arrangement we obtain a *Kelvin–Voigt element*. In this way a sudden increase of the force ($\mathbf{F}$), leads to an exponential elongation of the system which will contract in the same way if the force vanishes.

Real systems must be regarded as being made up of both Maxwell and Voigt elements in various combinations and with different properties (e.g., *Maxwell–Weichert models*). Correspondingly, complicated strain-relaxation graphs and complex kinetic behavior is to be expected. If the time constants of the viscous elements are large

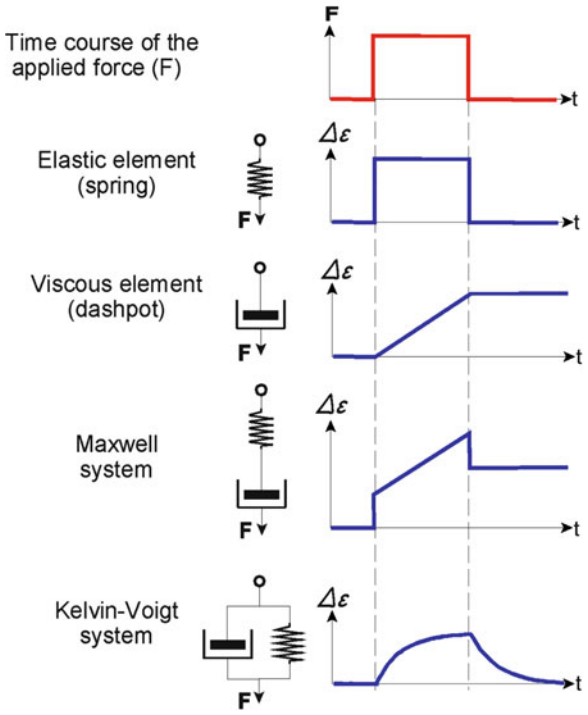

Fig. 3.56 Mechanical models to demonstrate the kinetics of elongation [$\Delta\varepsilon$(t)] of elastic, viscous, and viscoelastic elements after applying a rectangular function of force [**F**(t)](—)

in comparison to the length of the mechanical impulse, i.e., in the case of a short-term mechanical stress, viscoelastic systems can respond elastically.

Parallel to the strain function [$\varepsilon(t)$] as a result of an applied function of force [F(t)] (*isotonic tension*) as illustrated in Fig. 3.56, an examination of materials is also possible by applying a definite strain function and measuring the resulting stress in time dependence [$\sigma(t)$] (*isometric tension*). In this case the Maxwell element shows an exponential decline of stress as a result of viscous elongation of the dashpot relaxing the spring.

The schematic graphs of Fig. 3.56 with their different combined mechanical elements are similar to the electrical RC circuits in Figs. 3.39 and 3.43. In fact there are some similarities in their kinetic treatment. In analogy to the electrical impedance which we discussed in Sect. 3.5.3, a kind of mechanical impedance can be formulated as the response of a viscoelastic system to time-varying mechanical forces. Measurements of this kind give information on basic mechanical properties of biological tissue.

The modulus of elasticity (Y in Eq. 3.226) of various materials differs in several degrees of magnitude. In contrast to the value for steel, which is about $2{\cdot}10^5$ MPa, there are only some tenth to hundreds of MPa for various kinds of rubber. This marks also the range for different elastic properties of biological materials. Resilin, the most elastic animal protein from locust tendons shows a Young's modulus even

below one MPa. For bone, values of several hundreds of MPa are determined. Wood, stressed along the grain arrives at an elasticity of up to 10^4 MPa.

For the molecular basis of deformations, two mechanisms must be considered: In the case of *steel elasticity*, which occurs in most crystalline materials, the elongation leads to changes of the atomic spacings in the crystal lattice. The atoms are forced to move from their equilibrium position of minimal internal energy to a higher energy level. The free energy that is stored up in the strained state is released during relaxation. This is the reason for the high elasticity modules and also for the only short amount of elastic strain of these materials near to the rupture point.

In the case of so-called *rubber elasticity*, which is typical for macromolecular materials, the deformation of the body is the result of molecular deformations. We discussed this type of elasticity in Sect. 2.1.6 (Fig. 2.11). The applied strain leads to a partial deconvolution and orientation of the molecules and as a result, to a reduction of their entropy. The subsequent relaxation is the result of the increase in entropy back to its maximum according to the second law of thermodynamics. This is the reason why this type of stretching mechanism is also called *entropy elasticity*. The characteristic differences to steel elasticity are the large amount of possible elongation, and the strong temperature dependence of rubber elastic materials.

A number of structure proteins like for example collagen, elastin, or resilin are responsible for the rubber elastic properties of tendons and ligaments. These substances also play an important role in the storage of mechanical energy in some periodic or suddenly occurring movements. Although these processes also occur in mammals, they are best investigated in insect jumping and flying. In the case of locust jumping for example, the muscles apply a tension to a system of tendons which takes up the energy and subsequently releases it, by means of a trigger mechanism to achieve an increase in power. In this way a power, i.e., a transformed energy per time can be generated, which is impossible to obtain directly from the muscle. In fact, this is the same mechanism that we use for archery. To carry out a jump, a grasshopper for example, needs a specific power of about 5 kW per kg muscle. This would exceed the maximal output of a muscle by ten times. Similarly, during flight a grasshopper stores about 20–30% of the energy of the oscillating movements of the wings using passive elastic elements.

The following characteristics of the viscoelastic behavior of biological systems, such as cells, tissues, and organs should be noted:

The occurrence of different regions of elasticity in the stress–strain diagram: Frequently cells and elastic fibers in a tissue are interconnected to each other forming a network. In this case the viscoelastic properties of this system result not only in the intrinsic properties of these fibers, but first of all in the construction of this network, and in the viscosity of the fluid within it. Such a network can be stretched easily as long as it is deformable as a whole. This is the first region in the stress–strain diagram with low Young's modulus. If the components of the network eventually are fully oriented by the applied stress, a further strain is possible only by stretching the fibers themselves. This means that the Young's modulus increases suddenly. The resulting stress–strain diagram therefore, does not indicate a

flattening of the curve as in Fig. 3.55, but on the contrary it becomes steeper at a certain degree of strain. This sort of behavior for example, occurs in the walls of blood vessels, where the situation becomes even more complicated due to the activity of the smooth muscles.

The regulation of the elasticity behavior: The viscoelastic properties of the network as described above can be controlled by biological processes. This for example is possible by loosening and fastening the points of connections between the components. The result would be a sort of viscous prolongation of the body. Another way to control the viscoelastic behavior is the alteration of the water content of the tissue. Changing the cell volume or the intercellular space in the network would change its elasticity modulus. This sort of control is best investigated for the case of viscoelastic properties of the uterine cervix at various periods of pregnancy.

Mechanical anisotropy: In many biological systems the elasticity modulus depends on the orientation of the applied stress. This property is best investigated in bones. Depending on the angle of measurement, Young's modulus in bones can vary by a factor of two. The reason for this is the particular structure of the bone. The struts of the cancellous bone, the so-called *trabeculae* are oriented according to its loading in vivo. They are oriented according to the trajectories of pressure and tension of the bone in the skeleton. This is the result of self-orientation and adaption, which was first investigated in 1892 by the German anatomist Julius Wolff formulating the law of bone remodeling (*Gesetz der Transformation der Knochen*).

The special mechanical properties of the cell membrane have been discussed in detail in Sect. 2.3.4 (Fig. 2.40). A lipid membrane can be considered as a two-dimensional crystal. The head groups of the phospholipids show similar behavior to the atoms in the three-dimensional crystal of steel. Therefore, they are showing a kind of steel elastic behavior in the plane. The elasticity modulus of these membranes is high and their limit of rupture low. The membrane proteins with their rubber elastic properties have no significant influence on this behavior. In contrast to technical materials like rubber sheets, the biological membrane can easily be deformed in an isoplanar fashion, but cannot resist expansion. This property is important for the dynamics of cell shape and for cell swelling. So for example swelling of human erythrocytes can easily occur only by deformation up to the volume of a perfect sphere. The swelling of lymphocytes or other cells is possible thanks to smoothing of the membrane.

Further Reading

Viscoelasticity in general: Hosford 2009, viscoelastic properties of biomaterials: Oomens 2009, Skalak and Chien 1987, the historical paper: Wolff 1986.

3.6.4 The Biomechanics of the Human Body

Investigations of biomechanical conditions of the human body, its movement, its carriage under conditions of loading, etc. are important tasks in orthopedics, surgery, sports, and occupational safety. What kinds of loads of the muscles and joints result from a particular loading and a particular movement? How can pathologic deviations be cured by surgical operations? How can diseased bones and joints be replaced by artificial materials? In this case immunological tolerance must be realized as well as the adaption of the applied material to the viscoelastic properties of the living bone the artificial joints will be connected to.

Recently, greater effort has been directed toward the construction of mathematical models of the locomotor system including human walking movements. With computer simulation programs, hopefully surgical corrections can be performed with optimal success. Of course, in this context sports efforts should also be mentioned which optimize various techniques of jumping, running, etc.

In contrast to the structure of plants which can be modeled statically by systems of flectionally and torsionally loaded bonded structures, for animals and human systems dynamic body models are required. They are composed of combinations of elements that are stable towards pressure and bending with tendons and muscles as elements of tensile stability and contraction. This is illustrated in Fig. 3.57. The head of the horse for example, is held by a system of elastin tendons, the

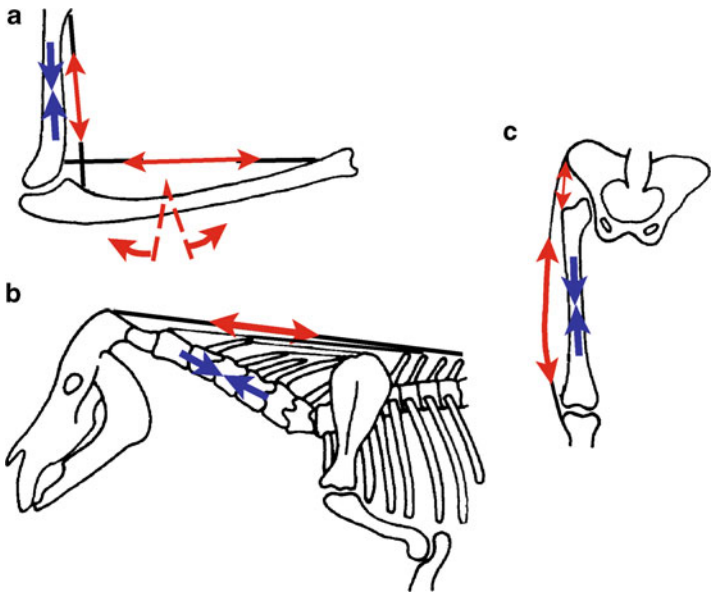

Fig. 3.57 Stabilization of the human forearm (**a**) and thigh (**b**), as well as the head of a horse (**c**) by the combination of bending and compressing stable bones, and tensioning stable muscles and tendons (After Glaser 1989)

ligamentum nuchae, as the tensile element and the cervical part of the spinal column as the component in compression. In the abdominal region of quadrupeds the compression element is located dorsally and the tension element is located ventrally. The carriage of the body is stabilized by tendons and it is maintained in an upright position without an additional supply of energy.

During evolution, the systems of combined elements which are stable against compression, together with those for tension, have been developed toward optimal use of muscle force, as well as toward maximal stability of the supporting bones. The attachment point of the muscles and tendons determines the vectors of tension and compression in the bone. We already mentioned in the previous section that this induces the oriented growth of the struts in the cancellous bone (Wolff's law).

Figure 3.58 shows an example of optimization of the shape of a human forearm. The bending of the bone and the shift of the muscle attachment have led to a significant reduction of the bending force. Similar principles of optimization can also be found in other parts of the muscle-skeleton apparatus.

The bending or the torque of a body can be attributed to stretching and compression of the material. As a measure of the bending, the *radius of bending* (R), or its reciprocal, the *curvature of bending* ($K = 1/R$) is used. If a homogeneous beam or bar bends, a compression of the concave side and a stretching of the convex side occurs. Between these two regions there must be a neutral plane that is neither compressed nor stretched.

Let us consider a section of a bar that is bent between two points that subtend an angle α at the center of curvature (Fig. 3.59). Let R be the radius of the curvature

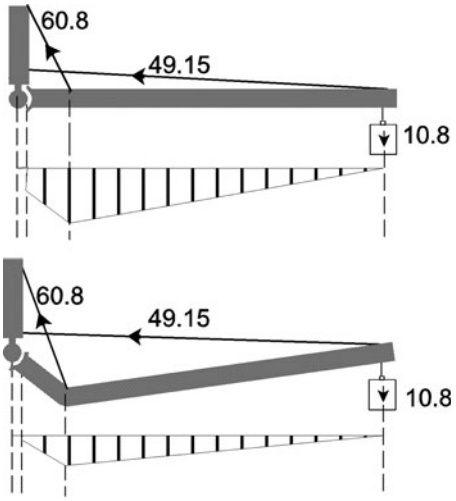

Fig. 3.58 Two steps for the optimization of the bending load of the forearm. The two muscles which held the forearm bone must generate forces of 60.8 N and 49.15 N, respectively, to compensate the loading force of 10.8 N at its end. These muscle forces are identical in both cases. The bending force of the bone, expressed by the graphs below, however, are quite different. Compare this scheme with the real situation in Fig. 3.57a (Redrawn after Pauwels 1980)

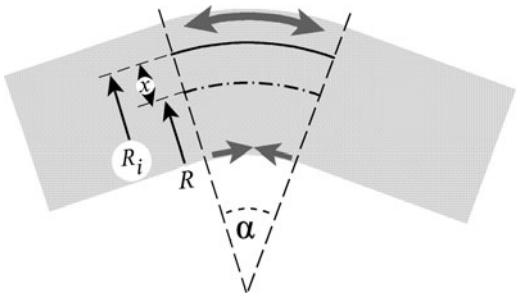

Fig. 3.59 The bending of a bar

measured to the neutral plane. If the angle is expressed in radiants then the length [$l(x)$] of the section at the distance (x) from the neutral plane is:

$$l(x) = \alpha\, R_i = \alpha(R + x) \tag{3.227}$$

If the length of the section along the neutral plane (at: $x_n = 0$; $R_n = R$) is l_n then the strain (ε) of any plane, parallel to the neutral plane can be calculated by Eq. 3.225:

$$\varepsilon(x) = \frac{\Delta l}{l} = \frac{l(x) - l_n}{l_n} = \frac{\alpha R_1 - \alpha R}{\alpha R} = \frac{x}{R} \tag{3.228}$$

As the distance x is measured from the neutral plane, ε can have both positive and negative values. Negative strain in this context means compression of the material and occurs when $x < 0$.

The combination of Hooke's law (Eq. 3.226) and the above relation enables the stress to be calculated:

$$\sigma(x) = Y\frac{x}{R} \tag{3.229}$$

The differential of moment of force (dM) is used to find the force which is necessary to bend the bar. It is calculated from the product of the force ($\mathbf{F}$) and the leverage distance from the neutral plane (x), and it is also related to the differential of an area (dA):

$$\mathrm{d}M = x \quad \mathrm{d}\mathbf{F} = x\,\sigma \quad \mathrm{d}A = \tfrac{x^2 Y}{R}\mathrm{d}A \tag{3.230}$$

The bending moment of the bar is obtained by integration of this equation. If the bar is homogeneous, the modulus of elasticity (Y) is not a function of x.

$$M = \frac{Y}{R}\int x^2\ \mathrm{d}A = \frac{Y}{R} I_A \tag{3.231}$$

In this equation the integral expression has been replaced by the *area moment of inertia* (I_A). This is a measure of the bending resistance of a bar made out of a material having a modulus of elasticity (Y). It can be seen that for a given bending moment, the smaller the value of I_A the more the bar bends. This means that if the bending moment does not change, and the modulus of elasticity remains constant, then the second moment of area is directly proportional to the radius of curvature.

It is easily understood from everyday experience that the area moment of inertia of a bar depends on the plane in which it is bent. A bar having a flat section bends more easily across the flat then it does on edge. This is determined by the position of the neutral plane. It passes through the center of gravity of a cross-section of the bar, and is always perpendicular to the radius of curvature.

For bars with particular geometrical profiles analytical expressions for I_A are derived. Comparing these values for a compact cylindrical rod with a radius r, with that of a tube with an inner radius r_1, and an outer radius r_2, the advantage of a structure made up of hollow tubes, is evident (Fig. 3.60). This optimizing principle has occurred in the construction of bones and some plant stems.

In the case of geometrically nondefinable structures the area moment of inertia can be determined by iterative methods. First it is necessary to find the position of the neutral plane. This is perpendicular to the bending radius and located at the center of gravity of the profile. For structures of homogeneous density the center of gravity can easily be found using a cardboard model of the cross-section. The center of gravity is the crossover point of all lines which can be drawn in a vertical direction from various points of suspension (Fig. 3.61).

After the position of the neutral plane has been fixed, the cross-section can be subdivided into rectangular areas (Fig. 3.62). The following approximation can be obtained from the definition of the second moment of area:

$$I_A \approx \sum_{1=1}^{n} x_i^2 \, \Delta A_i \tag{3.232}$$

The individual rectangular areas (ΔA_i) are calculated, the distance of their centers from the neutral plane (x_i) are measured, and the products of these two

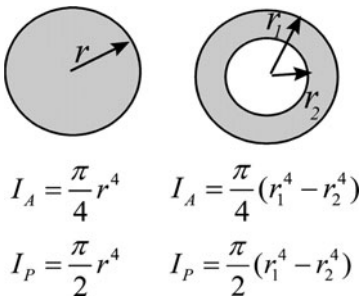

Fig. 3.60 Area moment of inertia (I_A) and polar moment of inertia (I_P) for a solid shaft and a tube

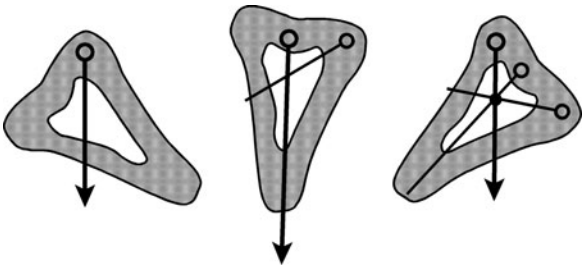

Fig. 3.61 Determination of the center of gravity of a tube of arbitrary cross-section as the crossover point of perpendicular lines drawn from the points of suspension

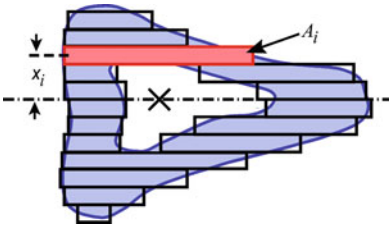

Fig. 3.62 Determination of the second moment of area of a tubular bone by a graphical method. The profile of the bone is approximated by many rectangles with individual areas A_i and distances from the neutral plane (–·–·–·–) x_i; X – center of gravity

measurements are summed as required by Eq. 3.232. The accuracy of this method will be increased if the areas are made as small as possible. The units of the second moment of area are m^4.

In a similar way to bending, the torsional deformation can be calculated. In this case not a neutral plane exists between the stretched and the compressed areas, but all regions are stressed as much as they are away from a central axis of gravity. A *polar moment of inertia* (I_P) can be formulated where the area elements (dA) are multiplied by the square of the radial distance (r) from the center of gravity, and subsequently, the products are summed. In an integral formulation this means:

$$I_P = \int r^2 \, \mathrm{d}A \tag{3.233}$$

In this way the polar moment of inertia of bars and tubes of definite geometrical profile can be calculated (Fig. 3.60).

These statements form the basis for calculating the stability of structural elements of plants, animals, and humans. However, it can be seen from the above equations that some simplifying assumptions have to be made. In particular it must be noted that the modulus of elasticity (Y) may vary with the position and also, as

illustrated by the properties of bone, with the direction of the applied force. If the viscoelastic properties are also taken into account, the calculations become even more complicated.

Further Reading

Biostatics of humans and animals: Fung 1993; Skalak and Chien 1987; biostatics of plants: Niklas 1992; Bone modeling: Carter and Beaupré 2001; Martínez-Reina et al. 2009.

3.7 Biomechanics of Fluid Behavior

Streaming of viscous fluids occurs at all levels of biological structure. There is water flow through pores of membranes, streaming of cytoplasm in plants, streaming of blood in the vessels, and finally flow of water and air around animals, i.e. the problems of flying and swimming. In this section we will concentrate on some medically important problems of hemodynamics, or hemorheology. But as this is a classical problem of biomechanics, some basic aspects of flying and swimming will be included.

3.7.1 Laminar and Turbulent Flows

When a liquid flows along the surface of a thin plate, a flow profile is formed over this boundary which changes its character as the distance (l) from the leading edge of the plate (at: $l = 0$) increases. This effect is illustrated in Fig. 3.63.

Directly near the surface of the plate there is a trapped layer, i.e., a fixed, nonmoving film of air or liquid (not marked in Fig. 3.63). At increasing distance from the plate (z), the velocity of the flow increases. Near the edge of the plate, up to a particular critical length (l) there is a region of laminar flow, where fluid layers (lat.: *lamina*) slide over one another in a parallel direction. This boundary layer is characterized by a velocity gradient ($\mathrm{d}\mathbf{v}/\mathrm{d}z \neq 0$) perpendicular to the surface of the plate (see Fig. 3.64). There is no sharp transition between this layer and the region of unaffected bulk flow ($\mathrm{d}\mathbf{v}/\mathrm{d}z = 0$, $\mathbf{v} = \mathbf{v}_\infty$). Usually the thickness of this laminar boundary layer is defined by the distance $z = \delta_L$, when $\mathbf{v} = 0.99\mathbf{v}_\infty$. As can be seen in Figs. 3.63, and 3.64, δ_L gets larger as the distance from the leading edge of the plate increases.

The region of laminar flow in relation to the distance from the edge of the plate decreases at increasing bulk velocity ($\mathbf{v}_\infty$). With increasing distance (l) from the edge, the boundary layer gets thicker and becomes increasingly unstable. At a critical point, this leads to spontaneous appearance of turbulences. As already explained (Sect. 3.1.3, Fig. 3.3) turbulent flow occurs if the streaming processes becomes nonlinear, i.e., if the linear flux-force relation (Eq. 3.49) is no longer

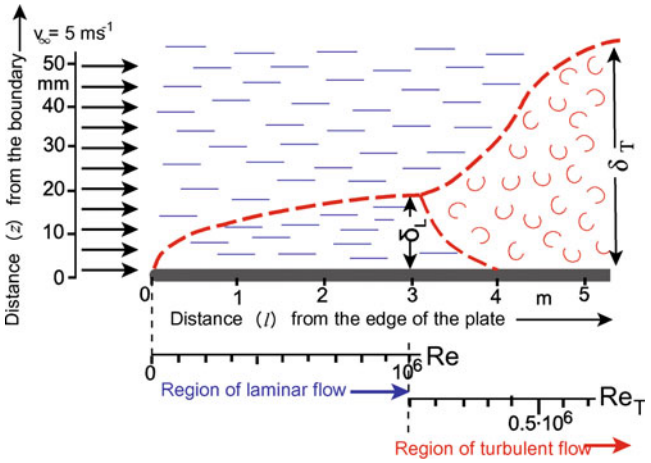

Fig. 3.63 Formation of laminar and turbulent flow profiles near a planar plate. As an example, the numbers correspond to a flow of air ($\nu = 1.5 \cdot 10^{-5}\,\mathrm{m^2\,s^{-1}}$) with a velocity $\mathbf{v}_\infty = 5\,\mathrm{ms^{-1}}$. The thickness of the laminar (δ_L) and the turbulent boundary layer (δ_T) are calculated according to the equations in Table 3.2

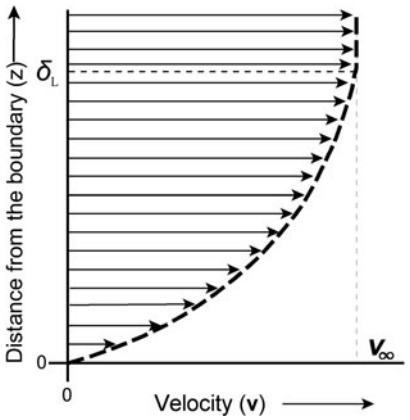

Fig. 3.64 Laminar velocity profile near a surface. The velocity vectors are depicted at various distances (z) from the surface; at $z = 0$, $\mathbf{v} = 0$; δ_L is the thickness of the laminar boundary layer

applicable. The transition from laminar to turbulent flow therefore is accompanied by an increase in friction and a substantial increase in the thickness of the boundary layer ($\delta_L < \delta_T$). The resulting whirls can be considered as a kind of dissipative structure.

In Table 3.2 some basic equations are listed to calculate parameters of laminar and turbulent flow. Partly, these equations are based on empirical observations, particularly those for turbulent flow.

It is not possible to determine exactly the critical point at which the transition from laminar to turbulent flow takes place. In fact, this is a stochastic process of

Table 3.2 Equations for parameters of laminar and turbulent flow near boundaries. Symbols (see also Fig. 3.63): **v** – velocity, z – distance from the boundary, ρ – density of the medium, η – viscosity, Re – Reynolds number; subscripts: 0 – at the boundary, ∞ – in the bulk phase

	Laminar flow (subscript L)	Turbulent flow (subscript T)
Sheer stress τ (z)	$\tau_0(1–z/\delta_L)$	
τ_0	$0,332\,\rho\, v_\infty^2 (\mathrm{Re}_L)^{-1/2}$	$0,023\,\rho\, v_\infty^2 (\mathrm{Re}_T)^{-1/5}$
Velocity $\mathbf{v}(z)$	$2\,\mathbf{v}_\infty/\delta_L(z–z^2/2\delta_L)$	$\mathbf{v}_\infty(z/\delta_T)^{1/7}$
Thickness δ	$5\,l_L(\mathrm{Re}_L)^{-1/2}$	$0{,}376\,l_T(\mathrm{Re}_T)^{-1/5}$
Force of surface friction $\mathbf{F}_0(l)$	$0,664\,\rho\, v_{\infty L}^{2l} (\mathrm{Re}_L)^{-1/2}$	$0,0366\,\rho\, v_\infty^{2l}{}_T (\mathrm{Re}_T)^{-1/5}$

state transition of the system that occurs as the result of increasing destabilization. The position of the critical point where the laminar flow abruptly transforms into a turbulent one can only be calculated as a matter of probability. It depends on the flow velocity (**v**), the viscosity (η), the density of the medium (ρ), and a characteristic streaming distance (l). These parameters are connected in the so-called *Reynolds number* (Re) which plays a crucial role in rheology:

$$\mathrm{Re} = \frac{l\,\mathbf{v}\,\rho}{\eta} = \frac{l\,\mathbf{v}}{\nu} \tag{3.234}$$

In this equation the kinematic viscosity ($\nu = \eta/\rho$) is used which has already been introduced in Sect. 3.6.1.

The Reynolds number is a typical parameter of the theory of similarity. Bodies of identical shape show identical flow behavior, for flow conditions with the same Reynolds number. This is independent of whether the body is large or small, or whether it is moving in water or air.

The critical Reynolds number characterizing the transition from laminar to turbulent streaming of a flow parallel to a flat surface, as illustrated in Fig. 3.63, is about $\mathrm{Re} = 10^6$. For a sphere with a flow around it, this transition already occurs at $\mathrm{Re} \approx 10^3$ (see Fig. 3.3). The critical Reynolds numbers of streamlined bodies with so-called laminar profiles are somewhere between these limits, depending on their exact shape.

Flow inside a cylindrical tube can be characterized in a similar way. In this case in Eq. 3.234 the radius of the tube is taken for the characteristic length l. The critical value of the Reynolds number for a flow in a tube is about 10^3. Turbulent flow means that the entire flow has become destabilized.

The following values for the kinematic viscosity can be used to calculate the Reynolds number for T = 291 K:

$$\nu_{\mathrm{Water}} = 1.06 \cdot 10^{-6}\mathrm{m}^2\mathrm{s}^{-1}$$
$$\nu_{\mathrm{Air}} = 14.9 \cdot 10^{-6}\mathrm{m}^2\mathrm{s}^{-1}$$

Table 3.3 Characteristic Reynolds numbers of various moving organisms

	Characteristic length (l) m	Characteristic velocity ($\mathbf{v}$) (m.s^{-1})	Reynolds number (Re)
Paramecium caudatum	$2.1 \cdot 10^{-4}$	$1.1 \cdot 10^{-3}$	$1.8 \cdot 10^{-1}$
Mosquito (*Ceratopogonidea*)	$0.9 \cdot 10^{-2}$	$2.5 \cdot 10^{-1}$	$1.5 \cdot 10^{2}$
Chaffinch	$3.6 \cdot 10^{-2}$	$2.1 \cdot 10^{1}$	$5.4 \cdot 10^{4}$
Crane	$2.6 \cdot 10^{-1}$	$2.8 \cdot 10^{1}$	$5.0 \cdot 10^{5}$
Water bug (*Dytiscus*)	$3.0 \cdot 10^{-2}$	$3.0 \cdot 10^{-1}$	$8.4 \cdot 10^{3}$
Stickleback (marine)	$1.0 \cdot 10^{-1}$	$7.2 \cdot 10^{-1}$	$5.5 \cdot 10^{4}$
Shark	$1.5 \cdot 10^{0}$	$5.2 \cdot 10^{0}$	$6.1 \cdot 10^{6}$
Dolphin (*Stenella* spec.)	$2.1 \cdot 10^{0}$	$9.3 \cdot 10^{0}$	$1.5 \cdot 10^{7}$
Blue whale	$3.3 \cdot 10^{1}$	$1.0 \cdot 10^{1}$	$2.6 \cdot 10^{8}$

If these values are substituted into Eq. 3.234 it is seen that, in contrast to streaming air, as shown in Fig. 3.63, turbulence in the case of water would occur not at $l = 3$ m, but already at $l = 0.2$ m from the leading edge. A comparison of flying and swimming objects can only be made under conditions of equal Reynolds numbers. It is quite senseless simply to relate velocities to the length of an object (which unhappily is often done in popular scientific publications, much to the amazement of the readers!).

Table 3.3 shows some typical Reynolds numbers for the movement of quite different organisms. In Sect. 3.7.3 we will show that, although they may possibly have optimal laminar flow shapes, large, fast-swimming fishes and aquatic animals exceed the critical Reynolds number for laminar flow.

3.7.2 *Biomechanics of Blood Circulation*

The biophysical properties of blood flow are quite complicated, even if all the complex physiological and biochemical control mechanisms are at first neglected. There are at least two features in which blood flow in general differs from the movement of a normal fluid through a tube, even having non-Newtonian properties. On the one hand, the elasticity of the blood vessels must be considered, changing their diameter as a function of the internal pressure, and on the other, it must be taken into account that blood is a suspension of flexible cells, the mean diameter of which may be of the same order as the smallest capillaries through which they must flow. The viscosity of the blood is a function of the cell concentration, the so-called *hematocrit*, which itself appears to be a function of the shear conditions in the vessel (see also Sect. 3.6.2). Furthermore, the pulsed character of blood flow must be considered and the particular aspects of bifurcations of the vessels.

Let us first consider some basic equations of fluid mechanics. Laminar flow through a tube may be thought of as the mutual movement of concentric hollow

cylinders. In this case each of these cylinders with a radius r, and a thickness dr experience a frictional force ($\mathbf{F}_F$) which is proportional to the velocity gradient (d$\mathbf{v}$/dr), to their surface area ($2\pi rl$), and to the viscosity of the fluid (η).

$$\mathbf{F}_F = 2\ \pi\ r\ l\ \eta \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}r} \tag{3.235}$$

The driving force ($\mathbf{F}_D$) behind such a flow can be obtained from the pressure difference (Δp) and the cross-sectional area of the cylinder (πr^2):

$$\mathbf{F}_D = \pi\ r^2\ \Delta p \tag{3.236}$$

In the case of stationary movement, both forces are balanced: $\mathbf{F}_F = \mathbf{F}_D$. Let the radius of the tube be r', and let us assume that there is a trapped boundary layer of liquid [$\mathbf{v}(r') = 0$] at the inner surface of the tube. Connecting Eqs. 3.235 and 3.236, and integrating them according to d$\mathbf{v}$ one obtains an equation for the velocity profile of this flow:

$$\mathbf{v}(r) = \frac{\Delta p}{4\,l\,\eta}(r'^2 - r^2) \tag{3.237}$$

Thus, $\mathbf{v}(r)$ is a parabolic function, whereas $\mathbf{v}_{\max}$ is the velocity at the center of the tube (see Figs. 3.51, and 3.66). In order to calculate the total volume flux ($\mathbf{J}_V$ in: $\mathrm{m^3\ s^{-1}}$) through the tube, the function $\mathbf{v}(r)$ must be integrated over the entire profile. This leads finally to the *Hagen–Poiseuille equation:*

$$\mathbf{J}_\mathrm{v} = \frac{\pi\ \Delta p\, r'^4}{8\,l\,\eta} \tag{3.238}$$

Thus, the flow through a tube is proportional to the fourth power of its radius. This is a very important aspect for the physiological control of local circulation. A slight widening or narrowing of the blood vessels causes large changes in the blood flow.

These fundamental laws of physical rheology must be considered just as a first approximation for what really happens in blood flow through the vessels. The following particularities must be taken into account, leading to an extension of these approaches:

– *Blood is a non-Newtonian fluid*, i.e., its viscosity (η) depends on the shear rate (γ) of the flow (see Sect. 3.6.2, Fig. 3.54). Integrating Eq. 3.235 in order to derive Eq. 3.237, we considered the viscosity as constant. This, obviously, is not acceptable in the case of blood. If, however, a particular function $\eta(\gamma)$ was applied, the integration would be more complicated, and the resulting velocity profile would not show a simple parabolic function.

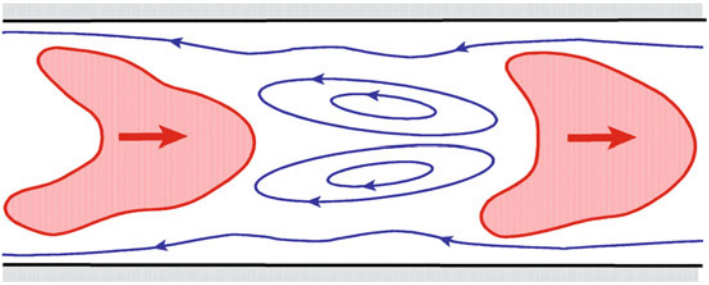

Fig. 3.65 Movement of deformed erythrocytes through a narrow capillary. The plasma, trapped in vortices between the cells is transported in the same direction together with them. Near the wall of the capillary plasma may move in the opposite direction (Modified after Talbot and Berger 1974)

– *Blood is not a homogeneous liquid, but a suspension of cells*. In capillaries, the diameters of which are of the same order of magnitude or even lower than the diameter of erythrocytes, the velocity profile of the plasma is determined by the moving cells which become strongly deformed in these narrow and branched vessels. This is a problem of *microrheology* of circulation (see Fig. 3.65). This particular streaming profile, especially the microscopic vortices in fact optimize gas exchange between tissue and erythrocytes.

  In large vessels the so-called *Fahraeus–Lindqvist effect* occurs. This leads the erythrocytes to concentrate in regions of minimal shear stress, namely in the center of the vessel. This means that the viscosity of the blood which we found to depend on the hematocrit (Sect. 3.6.1) increases in this region, but decreases near the wall of the vessel. This leads to a lowering of the streaming resistance of the total blood flow. Conversely of course, the streaming profile is changed dramatically. The parabola becomes flattened at the center of the vessel and steeper near the walls. Furthermore, this effect leads to a redistribution of different sorts of blood cells. In fact, the intensity of the force, shifting the cells by the Fahraeus–Lindqvist effect into regions of lower shear stress, depends on their size. As a result smaller cells like blood platelets are not influenced as much by this effect as erythrocytes with a larger diameter. This leads to concentration of the platelets near the walls of the vessel, which appears to be helpful in the case of injury.

  The Fahraeus–Lindqvist effect can be understood as the result of the Prigogine principle of minimal entropy production, as described in Sect. 3.1.4. It is valid for linear approaches, thus also for laminar flow (see Sect. 3.6.1). Minimal entropy production, for the case of blood flow means that the cells should concentrate at locations of minimal frictional energy dissipation, namely at locations of minimal shear rate.
– *The diameter of the blood vessels differs along the system of circulation*. If a tube suddenly becomes narrow a so-called *entrance effect* occurs (Fig. 3.66). This means that first the velocity profile of the narrow part of the tube corresponds to that of the central part of the broad tube. Only after a certain

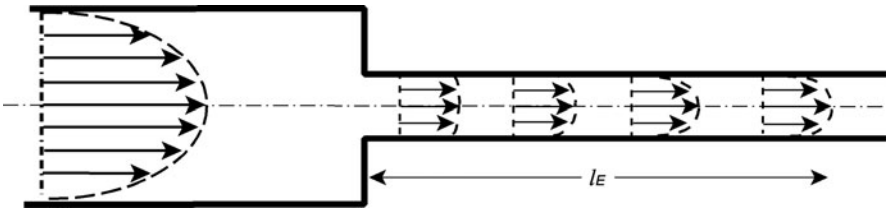

Fig. 3.66 Laminar flow in a tube: the parabolic velocity profile changes during a sudden narrowing of the tube's radius (entrance effect). Only at a distance l_E is a new parabolic profile established again

distance from the place of narrowing (l_E in Fig. 3.60), will a new profile be established. Usually this occurs at: $l_E = 0.06 \cdot r \cdot \mathrm{Re}$, where r is the radius of the narrow tube, and Re is the Reynolds number. This effect becomes important at the entrance of blood in the aorta. Furthermore, it occurs in the case of air flow in the lungs.

– *Blood flow is not stationary, but pulsed*. Therefore, the condition $\mathbf{F}_F = \mathbf{F}_D$ is no longer valid (see Eqs. 3.235 and 3.236). This fact is only important for arterial flow. The pulse waves of the heart partly become damped by the elasticity of the walls of the vessels, but they nevertheless proceed as oscillations of pressure and velocity of the blood flow up to the arterioles. One must differentiate between the true velocity of the blood streaming ($\mathbf{v}$) on one hand, and the velocity of pulse propagation ($\mathbf{v}_P$) on the other. The pulse propagation can be calculated using the *Moens–Korteweg equation*, which is derived for cylindrical tubes with thin walls:

$$\mathbf{v}_p = \sqrt{\frac{Y\ d}{2r\ \rho}} \tag{3.239}$$

In contrast to the flow rate ($\mathbf{v}$) (Eq. 3.237), the pulse propagation rate ($\mathbf{v}_P$) depends on the density (ρ) of the medium, but not on the viscosity (η). Conversely, the thickness (d), and the elasticity modulus (Y) of the wall of the vessels are included. This equation does not say anything about the damping of pulse waves.

To relate various pulsing flows, similar to the Reynolds number, another parameter of similarity (α) is introduced:

$$\alpha = r\sqrt{\frac{\omega\,\rho}{\eta}} \tag{3.240}$$

In this equation beside the density (ρ) and the viscosity (η), additionally the pulse frequency (ω) is used. This parameter α allows us to evaluate the relation between $\mathbf{v}_p$ and $\mathbf{v}$ under various streaming conditions. For small values of α the pulse propagation is faster than the velocity of the blood stream. At $\alpha \geq 3$, the pulse

propagation velocity becomes equal to the streaming velocity ($\mathbf{v}_P = \mathbf{v}$). In the aorta this parameter is higher than that limit ($\alpha = 15$). In the arteria femoralis this limit is reached ($\alpha = 3$). In these vessels therefore the pulses are propagated with the same velocity as the total flow rate of the blood. This consideration, however, does not include the fact that at points of branching of the vessels, reflections of the waves can occur.

To evaluate the elasticity of the blood vessels, let us start again with simple physical approaches. How is the radius of a tube changed by the strain of the tube wall? For the simplest case, Hooke's law (Eq. 3.226) can be applied, considering an elastic tube with a radius r, and a circumference of $2\pi r$. In this case the stress (σ) of the wall of the tube can be written as follows:

$$\sigma = Y \frac{\Delta r}{r} \tag{3.241}$$

We defined the stress (σ) as stretching force ($\mathbf{F}$) per area (A) of the stretched material (Eq. 3.224). Let the thickness of the wall be d, and its length l, the cross-sectional area of the stretched wall becomes $A = d \cdot l$. Therefore:

$$\sigma = \frac{\mathbf{F}}{ld} \tag{3.242}$$

Defining σ' as stretching force per length, gives:

$$\sigma' = \frac{\mathbf{F}}{l} = \sigma\, d \tag{3.243}$$

Now, we must know how large the stress (σ) would be in the wall of a tube with an internal pressure p. This can be calculated using the *Laplace equation* for a tube with radius r:

$$\sigma' = p\, r \tag{3.244}$$

Equations 3.241–3.244 enable us to derive a formula to calculate the increase of the radius of a tube (Δr) as a function of the internal pressure (p):

$$\Delta r = \frac{p r^2}{Y\, d} \tag{3.245}$$

Let us remember that we started with the assumption that the vessel wall shows a linear elastic behavior of the material according to Hooke's law. The relation between stress and strain therefore would correspond to Young's modulus (Eqs. 3.226 and 3.241). In fact the viscoelastic behavior of the wall of vessels is much more complicated and not at all linear. Furthermore, its strain is not only controlled by an interplay between various elastic materials, but is additionally

Table 3.4 Some rheological parameters of human blood circulation (Data from Talbot and Berger 1974)

Vessel	Average velocity m (s^{-1})	Diameter (m)	Average wall shear rate (s^{-1})	Reynolds number (Re)
Aorta	$4.8 \cdot 10^{-1}$	$2.5 \cdot 10^{-2}$	155	$3.4 \cdot 10^{3}$
Artery	$4.5 \cdot 10^{-1}$	$4 \cdot 10^{-3}$	900	$5 \cdot 10^{2}$
Arteriole	$5 \cdot 10^{-2}$	$5 \cdot 10^{-5}$	8,000	$7 \cdot 10^{-1}$
Capillary	$1 \cdot 10^{-3}$	$8 \cdot 10^{-6}$	1,000	$2 \cdot 10^{-3}$
Venule	$2 \cdot 10^{-3}$	$2 \cdot 10^{-5}$	800	$1 \cdot 10^{-2}$
Vein	$1 \cdot 10^{-1}$	$5 \cdot 10^{-3}$	160	$1.4 \cdot 10^{3}$
Vena cava	$3.8 \cdot 10^{-1}$	$3 \cdot 10^{-2}$	100	$3.3 \cdot 10^{3}$

regulated actively by smooth muscles. This interaction of passive and active properties of the vessel wall is of great importance for the regulation of blood flow.

All these calculations consider the blood flow in the system of circulation as laminar. Is this correct or are there turbulences in human blood flow? The numbers of Table 3.4 indicate that in fact, at least in large arteries critical Reynolds numbers can occur. We mentioned in Sect. 3.7.1 that the laminar flow in tubes becomes unstable in the case of Reynolds numbers near 1,000. This limit, however, is only correct for smooth, rigid and absolutely cylindrical tubes. In blood vessels, some factors help to stabilize the laminarity even beyond this limit, while others induce turbulent flow already at lower Reynolds numbers. Factors inducing turbulences are for example the branching of the vessels and inhomogeneities of their walls. These include arteriosclerotic alterations, or consequences of various surgical operations. In general, however, the system of blood circulation can be considered as biomechanically optimized.

Various techniques have been applied to investigate the properties of streaming blood. In medical diagnosis various ultrasound techniques are used to analyze flow properties of blood as well as viscoelastic properties of the vessels. Especially, the Doppler effect of ultrasound allows us to investigate blood flow, and even the flow profiles in large vessels. Special computer programs help to visualize these processes. This also allows us to check the blood flow and functions of the heart valves. Furthermore, investigations are undertaken analytically in tubes made of transparent materials which copy particular regions of human vessels (Fig. 3.67).

The model system of Fig. 3.67 shows special properties of flow behavior near the branching of vessels. The critical Reynolds numbers in these regions are lower than in unbranched regions. This depends on the angle of bifurcation (α in Fig. 3.67). For $\alpha = 180°$ the laminar flow already becomes critical at Re = 350. If the angle is only 165°, the critical point is at Re = 1,500. Additionally, the critical Reynolds number depends on the relation of the radius of these branches.

These few aspects of biophysics of blood rheology already show us how complicated this situation in vivo in fact is. This branch of biomechanics is developing quickly. This tendency is promoted by the fast progress in surgery of blood vessels on the one hand, and on the construction of artificial cardiac valves, of artificial hearts as well as of various systems of extra-corporal circulation on the other.

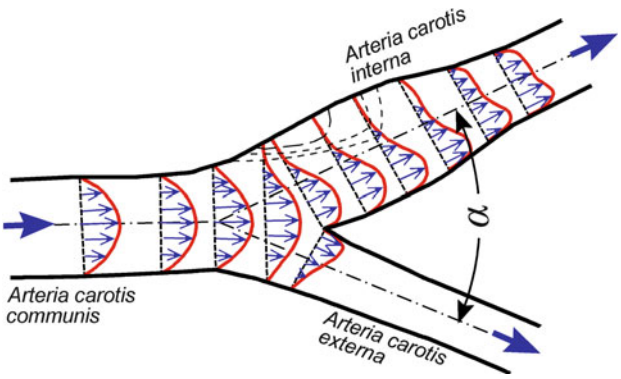

Fig. 3.67 Velocity profile in a model of the human carotis with 70% flow through the Arteria carotis interna. α – angle of bifurcation (After Schneck 1980 modified)

Further Reading

Bejan and Lorente 2008; Fung 1984, 1993, Skalak 1987, Waite and Fine 2007.

3.7.3 Swimming and Flying

The Reynolds number enables us to relate flow properties of moving objects in water with those of air. This allows us to relate the mechanisms of swimming to that of flying.

When a body moves relative to its surrounding medium then a *surface friction*, *or skin friction* occurs, and furthermore a drag on the surface of the body which is caused by its shape, the so-called *profile*, or *form drag*. The surface friction arises from phenomena which occur in the boundary layer and which have already been discussed in Sect. 3.7.1. Conversely, form drag arises because of the form (shape) of the object and is related to the volume of the surrounding medium that is displaced by the moving body. Objects with a larger cross-section perpendicular to their direction of movement will have a higher drag than thinner bodies.

Regions around the body therefore can become influenced to a much larger extent than those at the boundary layer in a fluid flowing parallel to a flat plate (Fig. 3.63). The critical Reynolds number, indicating that the laminar boundary layer becomes destabilized, therefore, is determined, to a great extent by the shape of the body. This explains why these two components of drag generation, namely the surface friction on the one hand, and the form drag on the other, cannot be treated separately from one another.

The boundary layer around a moving body is to a great extent influenced by local pressure differences. This arises from differences of the velocity (**v**) at different locations. Because of the law of conservation of energy, the sum of kinetic energy

of the moving medium ($½\rho\mathbf{v}^2V$), and the static energy of compression (pV) must be constant at all points of the space:

$$p + \frac{1}{2}\rho\ \mathbf{v}^2 = \text{const} \tag{3.246}$$

This constant is determined by the following conditions: at $\mathbf{v} = 0$ the hydrostatic pressure must be: $p = p_0$, whereas p_0 is the static pressure of the environment. This leads to the *Bernoulli equation*:

$$p = p_0 - \frac{1}{2}\ \rho\ \mathbf{v}^2 \tag{3.247}$$

The local pressure (p) acting on the surface of a body which moves in relation to a fluid, therefore results from the hydrostatic pressure (p_0) of the environment, lowered by the parameter $½\rho\mathbf{v}^2$. In addition to this pressure, however, an impact pressure must be considered affecting the points where the velocity vector of the streaming fluid is perpendicular to the body surface.

Figure 3.68 indicates the flow profile and the pressure distribution around a streamlined body, i.e., a body with minimal form drag. At points where the velocity of the flow is greatest, which in the diagram is indicated by maximal density of the lines of flow, i.e., near the middle of the body, there will be a maximum of negative pressure. At both ends the form drags dominate. These local pressures lead to forces which are directed always perpendicular to the surface of the body. At locations where this pressure becomes negative, i.e., at points where the forces are directed away from the surface, a disruption of the flow from the surface of the body can occur. This results in a wake of large eddies that is a region of turbulent flow which occupies a space much larger than that corresponding to the thickness of the turbulent boundary layers given in Fig. 3.63 and Table 3.2.

Considering this situation, there are three discrete qualities of flow pattern, obtained by increasing velocities or better, increasing Reynolds numbers:

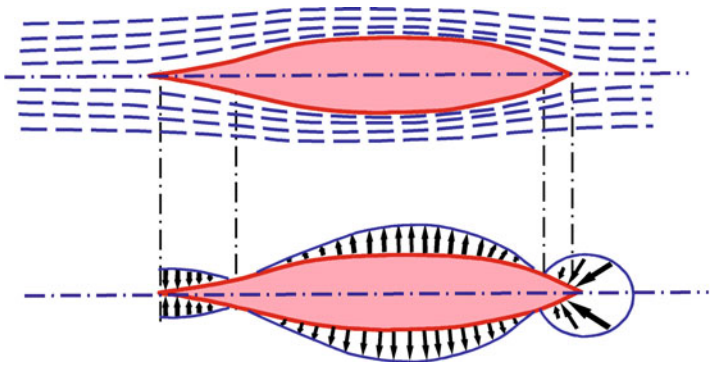

Fig. 3.68 Flow profile and pressure distribution around a moving streamlined body (Redrawn after Hertel 1963)

– Laminar flow around the body,
– Turbulent flow in a boundary layer,
– Turbulent, disrupted flow forming a wake.

The resistance to streaming around the body in these situations always increases stepwise.

In Table 3.3 Reynolds numbers are indicated for some swimming animals. The small animals, in spite of having unfavorable shapes, are in the region where the flow is always laminar, but for animals where the Reynolds number is in the order of 100 and higher, a hydrodynamic optimization of the shape of the body is required. In contrast to the sphere which allows laminar flow up to Re $\approx$ 1,000, the critical Reynolds number for a streamlined body, of course, is higher. The body shape of fast swimming animals really does appear to be optimal in this respect. Nevertheless, the Reynolds numbers of fast swimming fishes and dolphins, and also of many birds, lie in the supercritical range. In such cases various adaptions can be found that, at least, impede the disruption of the turbulent flow. This, for example, is achieved by the early induction of microturbulences at the surface of the body by particular surface structures such as feathers or scales.

Much has been written in papers on bionics about the specific adaptations of the dolphin which enable it to swim extremely fast. Apparently there is a viscoelastic damping layer of the skin and furthermore, the ability to induce folds at the body surface by muscular activity, both of which prevent the occurrence of latent instabilities in the flow.

The real friction of a swimming fish is difficult to measure. The simplest way would be to pull dead or anesthetized fish at a given velocity through water. The frictional resistance, obtained in this way, however, is so high that it would be impossible for the musculature of an actively swimming animal to overcome them. This leads to the conclusion: the fish is unable to swim! This circumstance was postulated in 1936 by Sir James Gray, and calculated particularly for the case of dolphins. The solution of *Gray's paradox* is as follows: In contrast to technical constructions, for example vessels containing two distinct elements, the driving component (screw), and the frictional component (body), in the case of living fishes or even dolphins, both elements are combined. A fish diminishes its friction during active swimming. The actively swimming animal has a lower drag than one that is towed passively through the water.

The part of the body that is involved in propulsion of fishes can be very different. Depending on the relative length and flexibility of the tail, three main types can be differentiated. The *anguilliform type* of movement, for example the eel, involves the whole body for propulsion. Most fishes show the *canrangiform type* of propulsion (Fig. 3.69). In this case tapering tails of medium length allow fast and dexterous swimming. They are able to accelerate quickly reaching a high speed after a short time. In the case of *ostraciiform type* of movement, named after the trunk of coffer fish, only the fins, like propellers propagate the fish. These fishes are only able to swim slowly.

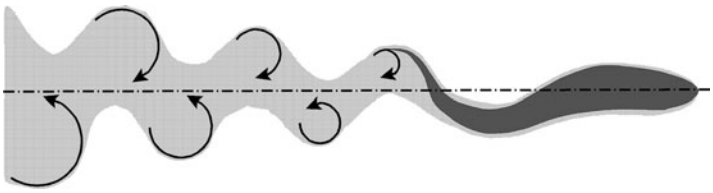

Fig. 3.69 Carangiform mechanism of propulsion of a fish with the vortex street behind

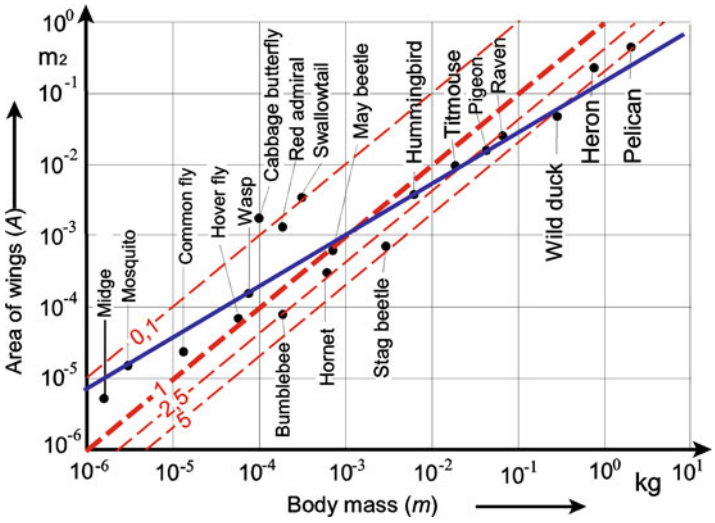

Fig. 3.70 The area of wings (A, in m^2) of various insects and birds as a function of body mass (m, in kg), follows approximately an allometric equation: $A = 0.11\ m^{2/3}$ (*blue line*). The *red lines* demonstrate the resulting wing loading (in $kg \cdot m^{-2}$) (Data from Hertel 1963)

In contrast to swimming, where we considered mechanisms of drag reduction, in the biomechanics of flying, lift is of central interest. First, we must distinguish between passive gliding and active flying. There are not only various insects, especially butterflies, birds, and mammals that can glide, but also a number of plant seeds and fruits. Many types of aerodynamic mechanisms are employed by these vegetable objects ranging from simple adaptations to reduce the rate of descent by means of rotating wings, for example the propeller seed of the maple tree, up to the stable gliding flying-wing of the seeds of the climbing pumpkin species *Zanonia macrocarpa*.

Active flying by animals is achieved through a most varied assortment of mechanisms. Those animals which can rise into the air without some initial help and then move forwards or backwards from this position can be considered as ideal fliers. This hover flight, however, requires power which dramatically increases with the wing loading, i.e., with the relation between body weight and wing area. Figure 3.70 shows that in fact the area of wings in various flying animals does

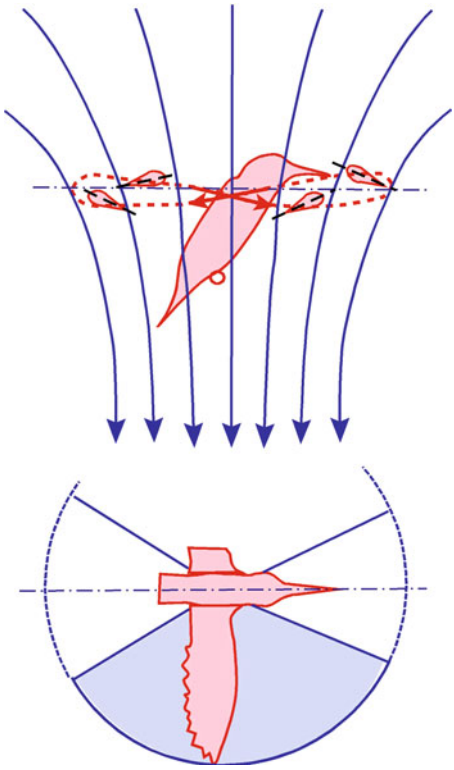

Fig. 3.71 Wing position and air flow of a hovering humming bird (Redrawn after Hertel 1963)

not increase in direct proportion to their body mass, but rather to the power of $\frac{2}{3}$. This is a common allometric relation which we will discuss in the next section (Sect. 3.8). The isometric lines (in Fig. 3.70: red lines, all with a power of 1) allow us to evaluate the real wing loading. The upper limit of the specific muscular performance is already reached in the case of humming birds. Larger birds and even large insects, like the stag beetle (see Fig. 3.70) are not able to hover on their own although some of them, like the kestrel can do so with the help of the wind.

Figure 3.71 shows schematically the wing movement of a hovering humming bird. The wings move horizontally, and each covers about a third of the horizontal area around the bird. The angle of attack is continuously adjusted so that lift is generated as the wing moves forwards, as well as backwards. In contrast to humming birds, the wings of flies and bees move in a vertical direction. The wings are also twisted but only the downstroke is used to generate lift. In the case of birds flying forward, the wings are moved up and down. The angle of attack is also regulated, but additionally the wings are bent or partly folded on the upstroke.

In contrast to the humming bird, all other birds require some assistance to get off the ground. Just as an aircraft must reach a take-off speed to ensure that the wings are generating sufficient lift, so must birds. This is achieved by running, jumping, or dropping. Nevertheless, the take-off speed of birds is remarkable low. This means

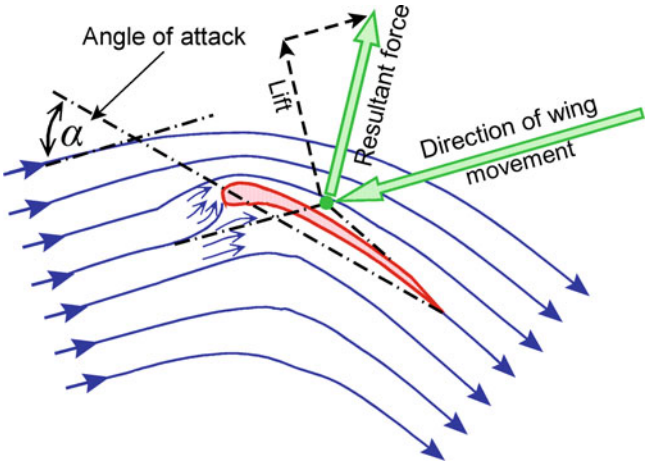

Fig. 3.72 Air stream and forces of bird flight during a stroke of the wing forwards and downwards (Redrawn after Hertel 1963)

that maximum lift at minimal Reynolds number is required. Figure 3.72 shows the force diagram for a wing that is being thrust forwards and downwards. If the incident flow remains constant ($\mathbf{v}_0$) then the lift generated by the wing increases with the degree of curvature of the wing section as well as with the angle of attack (α). The amount of increase of both these parameters however, is limited. If they exceed a critical value, the flow is disrupted, and turbulence develops. This will not only cause a loss of thrust but will at the same time destabilize the entire system. Such an occurrence would cause an airplane to crash. There are a number of biological adaptions preventing such a disaster and at the same time ensuring maximum lift at low speeds.

Further Reading

Ahlborn 2005; Alexander 2003; Azuma 1992; Leyton 1975; Videler 1993; Vogel 1994; Webb and Weihs 1983.

3.8 Allometric Considerations of Structure and Function

We already touched upon the problem of scaling of various processes in some earlier sections. So, the Reynolds number (Sects. 3.7.1, 3.7.3), and the α-parameter (Sect. 3.7.2) enable the mutual compatibility of streaming behavior of structures with various sizes. In fact, there exists a long list of dimensionless numbers of this kind which are used in technical engineering, allowing us to scale streaming, convection, heat conduction, and many other processes. But it is not only in engineering that it is necessary to compare structures and functions of systems

with different size, in fact, it is attracting more attention also in regard to various biological systems.

From the smallest organism to the largest, the size ranges through many orders of magnitude. Additionally, there are considerable size differences in the individual development of a single animal, as well as in different animals of the same species. To relate various biological functions with size, i.e., mass, is not only a general question of theoretical understanding of biological organization, but in some cases it is quite important for practical problems in medicine, sports, agriculture, etc.

For many biological variables a relationship can be written in the form of a so-called *allometric function*:

$$y = \alpha x^{\beta} \tag{3.248}$$

where α is the *allometric coefficient* and β the *allometric exponent*. The word *allometric* comes from the Greek and means "different measure." As a matter of fact, the relation is said to be allometric only if $\beta \neq 1$. Otherwise the relationship is linear and it is called not allometric, but *isometric*.

In general, we find three levels of scaling in biology: scaling relations within an individual organism during its growth (*ontogenetic allometry*), scaling relations among individuals of one species (*intraspecific allometry*), and finally scaling relations for individuals of different taxonomic or functional groups (*interspecific*, or *phylogenetic allometry*).

This problem in general involves the question of similarity, which was formulated already in Euclidean geometry. Archimedes, more than 2,000 years ago formulated that the surface of bodies with similar shapes grows in proportion to the square of their size, whereas their volume, correspondingly, with the cube. Galileo Galilei in the seventeenth century already speculated about the similarities of animals. He noted that a dog may be able to carry two or even three similar dogs on his back in contrast to a horse which hardly can do the same. He furthermore emphasized that the thickness of the bones in animals of different sizes are not simply proportional to the linear dimensions of their body.

Physiologists have for a long time formulated allometric relationships between body mass and various structural and functional properties of animals of various size. In the center of interest is metabolic rate as a function of size. This is an important parameter because it limits almost all biological processes at different levels of organization. In aerobic organisms, it is equivalent to oxygen consumption, and can be determined in this way. From a thermodynamic point of view, neglecting the storage of chemical energy, it must finally result in heat production, and in fact, it represents the entropy production which is expressed by the dissipation function $\mathbf{\Phi}$ (see Sect. 3.1.4, Eq. 3.64).

In 1883 the German physiologist Max Rubner formulated a *surface law of metabolism*. He reasoned: if an animal is n times as big as another, then its surface (A) should increase by n^2 and its volume or mass (m) by n^3, provided that the density (ρ) of all organisms is more or less the same. The relations $A \sim n^2$ and

$m \sim n^3$, can be transformed into $A \sim m^{2/3}$. Rubner considered that the metabolism of an animal will finally produce heat (dQ_M/dt). He supported these considerations with experimental data, derived from measurements of heat production and oxygen consumption in animals of the gamut of sizes. To keep the temperature of the organism constant, there must be a steady state, i.e., the heat must continuously be dissipated from the skin's surface. This leads to $\mathbf{\Phi} \sim A$, or $\mathbf{\Phi} \sim m^{2/3}$.

To determine these allometric parameters usually a log-log plot of the data is used. Taking the logarithms of both sides the above-mentioned generalized allometric equation will be transformed into:

$$\log y = \log \alpha + \beta \log x \tag{3.249}$$

Plotting ($\log y$) against ($\log x$) one obtains a straight line with the slope β. The parameters α and β therefore are the results of a linear fit to the log-transformed data.

In contrast to Rubner's data, another physiologist Max Kleiber in 1932 stressed that β was larger than Rubner postulated, namely $\beta = 3/4$. In this way the discussion in animal physiology about this allometric function ($\mathbf{\Phi} = \alpha\, m^{\beta}$) began and proceeds even today. The following questions arise: what is the real value of the allometric coefficient α and the allometric exponent β? Are these parameters really constant for mammals of different size, from the mouse to the elephant, or even in other animals like reptilians, or birds? Considering the large amount of papers published in recent years, and the hundreds of measurements on various animals, some researchers maintain Kleiber's proposition, while others support the $\beta = 2/3$ parameter of Max Rubner.

What may be the reason for this controversy? This should be clarified first, before further theoretical explanations are helpful. The parameter $\mathbf{\Phi}$, representing the metabolic rate of an organism, of course varies depending on its physiological condition. A so-called *basal metabolic rate* (BMR) is formulated, representing the minimum power which is required to maintain the tissues and essential life functions in nontorpid animals. Alternatively, the metabolic rate arrives at a maximum (MMR) when the animal is running at top speed. This can exceed the BMR by a factor of ten, and the allometric exponents for MMR data have been found to increase up to nearly 0.9. In fact, the measured metabolic rate is in any case between these two extremes and therefore variable. Furthermore, the metabolic rate strongly depends on the temperature of the environment and the corresponding adaption. Therefore, the allometric exponent may also depend on geographical regions and habitats.

In general, all biological variables have a measurement error and therefore deviate from expected curves. Moreover, it is difficult to standardize the measuring conditions for different animals. In general, the value of the exponent depends on the conditions under which the data are recorded. Finally, the parameters of the allometric functions to some extent depend also on the applied fitting procedure.

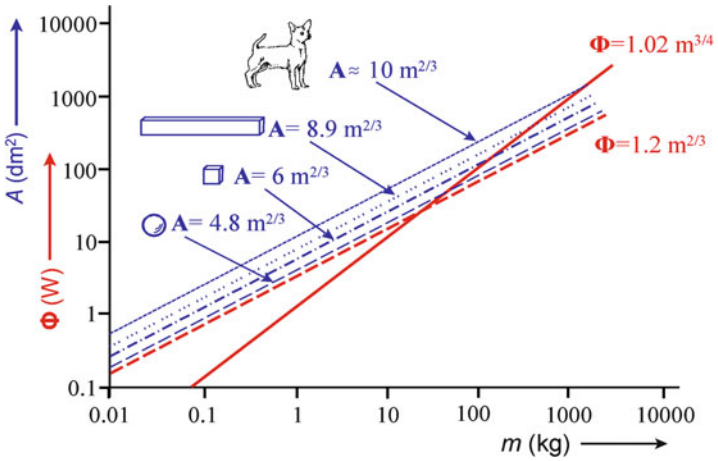

Fig. 3.73 Allometric functions of metabolism (*red*) according to Rubner's surface rule (*broken red line*: $\mathbf{\Phi} = 1.2\ m^{2/3}$), and Kleiber's rule (*full red line*: $\mathbf{\Phi} = 1.02\ m^{3/4}$), as well as simple surface mass relations (suppose: $\rho = 10^3\ kg\ m^{-3}$) of a sphere, a cube, a 1:10 prolonged cube, and that for a dog as a representative quadrupedal animal (*blue lines*)

Furthermore, the problem of geometrical and of course functional similarity of the animals should be considered. How can ectothermic animals relate to endothermic, herbivores to carnivores, birds to quadrupeds, etc. In fact, differences of the allometric parameters have been found in these special cases.

What are the real surface–volume relations of various bodies, and how do they influence the allometric function (Fig. 3.73)? Simple geometrical considerations show that the surface of a sphere grows with a function $4.84\ V^{2/3}$, that of a cube by $6\ V^{2/3}$. If one dimension of the cube is prolonged 10 times, the result would be $8.89\ V^{2/3}$. For mammals, factors between 9 and 11 have been estimated. This means that even for not fully geometrically similar animals, the surface-to-volume relation, or if the density (ρ) is constant, even the surface–mass relation may only differ in terms of the allometric coefficient, but not the allometric exponent.

Recently, a number of theories have been proposed to explain the allometric exponent of metabolism and its variation according to size and physiological condition. Instead of the body surface as considered by Rubner, mostly the transportation network of blood circulation was taken into account. So, for example a quarter-power law was derived, based mostly on geometry, particularly the hierarchical nature of circulatory networks. It seems however, that none of these models is complete, because the metabolic rate must account for all irreversible losses.

Allometric considerations are in fact not only centered on problems of metabolism. In Fig. 3.70 the area of wings of various insects and birds is plotted as a function of body mass. This also shows an allometric exponent of around 2/3. In fact, similar functions are considered in relation to a large number of structural and functional parameters such as size of legs, thickness of bones, dimensions of eggs, lifetime, heart frequency, speed of running, swimming, etc.

These parameters are always first estimated experimentally, followed by attempts at theoretical explanation. Sometimes simple relations are combined with measured allometric exponents. For example, the number of steps (n), a running animal must take per unit distance, inversely related to the length of the steps, and therefore inversely proportional to its length (l). This in turn can be related to the body mass:

$$n \sim \frac{1}{l} \sim m^{-1/3} \tag{3.250}$$

Furthermore, flat running can be related to the corresponding friction, whereas the climbing involved in running uphill means it must additionally account for the increase in potential energy. These types of speculation are usually connected again to metabolic and structural factors, deriving various dimensionless parameters.

Further Reading

Ahlborn 2005; Bejan and Marden 2009; Da Silva et al. 2006; Schmidt-Nielsen 1999; West and Brown 2004.