Chapter 3
Molecular Biology

Abstract Genetic information is passed with high accuracy from the parental organism to the offspring and its expression governs the biochemical and physiological tasks of the cell. Although different types of cells exist and are shaped by development to fill different physiological niches, all cells have fundamental similarities and share common principles of organization and biochemical activities. This chapter gives an overview of general principles of the storage and flow of genetic information. It aims to summarize and describe in a broadly approachable way, from the point of view of molecular biology, some general terms, mechanisms and processes used as a base for the molecular computing in the subsequent chapters.

3.1 DNA

In the majority of living organisms the genetic information is stored in the desoxyribonucleic acid (DNA), molecules that govern the development and functions of the organisms. The high accuracy of duplication and transmission of the DNA is determined by its structural features and the unique fidelity of the proteins participating in this process.

3.1.1 Molecular Structure

DNA is composed of four nucleotides, also called *bases:* adenosine (`A`), cytidine (`C`), guanosine (`G`), and thymidine (`T`), each of which consists of a phosphate group, a sugar (deoxyribose), and a nucleobase (pyrimidine – thymine and cytosine, or purine – adenine and guanine). The nucleotides are covalently linked through the sugar (deoxyribose) and phosphate residue and form the backbone of one DNA strand (Fig. 3.1). These two different elements

Z. Ignatova et al., *DNA Computing Models*,
DOI: 10.1007/978-0-387-73637-2_3,

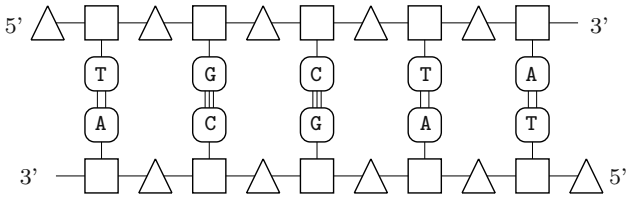

Fig. 3.1 Schematic overview of the DNA structure. The phosphate group is shown as a triangle, the sugar component is depicted as a square and together they form the backbone. The double helix is stabilized by hydrogen bonds between `A` and `T` (two hydrogen bonds) and `G` and `C` (three hydrogen bonds).

(sugar and the phosphate group) alternate in the backbone and determine the *directionality* of the DNA: the end with the exposed hydroxyl group of the deoxyribose is known as the 3' end; the other end with the phosphate group is termed the 5' end.

Two single DNA strands assemble into a double-stranded DNA molecule, which is stabilized by hydrogen bonds between the nucleotides. The chemical structure of the bases allows an efficient formation of hydrogen bonds only between `A` and `T` or `G` and `C`; this determines the *complementarity principle*, also known as *Watson-Crick base-pairing* of the DNA double helix. The `A` and `T` base pair aligns through a double hydrogen bond and the `G` and `C` pair glues with a triple hydrogen bond, which is the reason for the higher stability of the `G`–`C` *Watson-Crick base pair* over the `A`–`T` *Watson-Crick base pair*. The overall stability of the DNA molecule increases with the increase of the proportion of the `G`–`C` base pairs. The two single DNA strands are complementarily aligned in a reverse direction: the one, called also a *leading strand*, has a 5' to 3' orientation, whereas the complementary strand, called *lagging strand*, is in the reverse 3' to 5' orientation (Fig. 3.1).

In aqueous solution the two single strands wind in an anti-parallel manner around the common axis and form a twisted right-handed double helix with a diameter of about 20 Å. The planes of the bases are nearly perpendicular to the helix axis and each turn accommodates 10 bases. The wrapping of the two strands around each other leads to a formation of two grooves: the major is 22 Å wide and the minor is 12 Å wide. This structure is known as B-DNA and represents the general form of the DNA within the living cells. Alternatively the DNA double helix can adopt several other conformations (e.g., A-DNA and Z-DNA), which differ from the B-form in their dimensions and geometry. Unlike the A-DNA which is also right-handed, the Z-DNA is left-handed and the major and minor grooves show differences in width. The propensity to adopt one of these alternative conformations depends on the sequence of the polynucleotide chain and the solution conditions (e.g., concentrations of metal ions, polyamines). The hybrid pairing of DNA and RNA strands has under physiological conditions an A-form like conformation, while the Z-form

is believed to occur during transcription of the DNA, providing a torsional relief for the DNA double helix.

The concept of the DNA complementarity is crucial for its functionality and activity. The base-pairing can be reversibly broken, which is essential for the DNA replication. The non-covalent forces that stabilize the DNA double helix can be completely disrupted by heating. The collapse of the native structure and the dissociation of the double helix into two single strands is called *denaturation* (Fig. 3.2). Under slow decrease in temperature, the correct base pairing can be established again and the DNA renatures. The process of binding of two single strands and the formation of a double strand is known as *annealing* or *hybridization*. The annealing conditions need to be established by a slow change of temperature, as a rapid decrease in temperature forces a fast renaturation and results in both intramolecular (within one strand) and intermolecular (within different strands) base-pairing (Fig. 3.2). Complementary stretches within one single strand that are in close proximity can re-associate to partial double-stranded intramolecular structures, known as *foldback structures*.

The DNA double helix is very stable; the entire network of hydrogen bonds and hydrophobic interactions between the bases is responsible for its global stability. Nevertheless, each single hydrogen bond is weak and short stretches from the double-stranded DNA can even be opened at physiological temperature with the help of *initiation proteins*. Each strand in the DNA serves as a template for the replication machinery, with the DNA polymerase

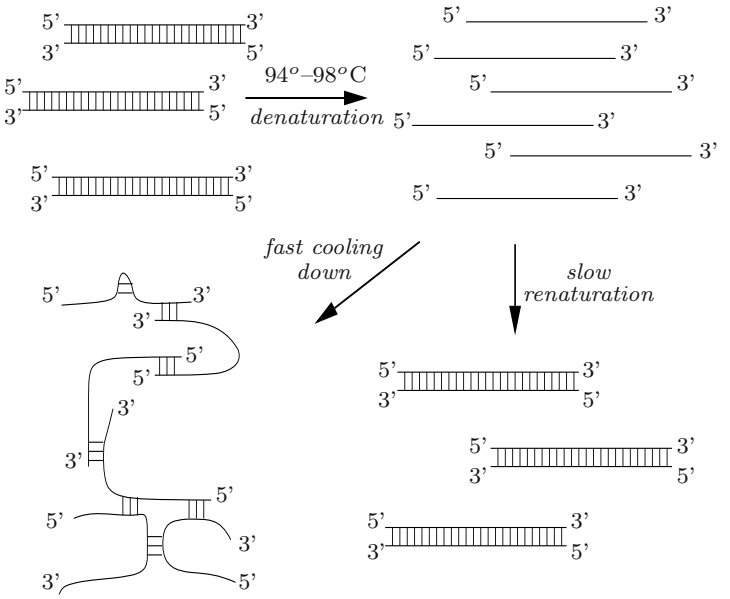

Fig. 3.2 Denaturation and renaturation of DNA.

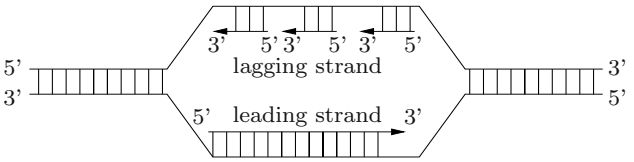

Fig. 3.3 Replication of DNA.

as a major player, replicating them in a complementary manner (Fig. 3.3). The DNA replication is asymmetric and DNA polymerase elongates in the 5' to 3' direction only. The opposite DNA strand is discontinuously synthesized again in the 5' to 3' direction as small fragments, called *Okazaki fragments*. The Okazaki fragments are further covalently joined by DNA ligase. In each replication cycle the double-stranded DNA template is replicated into two identical copies.

3.1.2 Manipulation of DNA

In DNA computing, DNA is utilized as a substrate for storing information. Depending on the model of DNA computation, information is stored in the form of single-stranded DNA and/or double-stranded DNA molecules. This stored information can be manipulated by enzymes. One class of enzymes, *restriction endonucleases*, recognizes a specific short sequence of DNA, called *restriction site*, and cuts the covalent bonds between the adjacent nucleotides (Fig. 3.4). Restriction fragments are generated with either cohesive or sticky ends or blunt ends.

DNA ligase covalently links the 3' hydroxil end of one nucleotide with 5' phosphate end of another, thus repairing backbone breaks (Fig. 3.5). The *exonucleases* are enzymes that hydrolyze phosphodiester bonds from either the 3' or 5' terminus of single-stranded DNA or double-stranded DNA molecules and remove residues one at a time. Endonucleases can cut individual covalent bonds within the DNA molecules, generating discrete fragments.

BamHI	SmaI
$5' - \texttt{G}\vert\texttt{GATCC} - 3'$	$5' - \texttt{CCC}\vert\texttt{GGG} - 3'$
$3' - \texttt{CCTAG}\vert\texttt{G} - 5'$	$3' - \texttt{GGG}\vert\texttt{CCC} - 5'$

Fig. 3.4 Restriction sites of BamHI and SmaI. The restriction enzymes recognize palindrome sequences with a two-fold rotational symmetry.

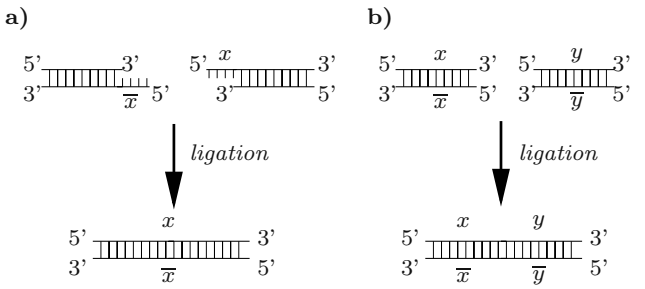

Fig. 3.5 Ligase connects **a** sticky or **b** blunt ends of double-stranded DNA.

Polymerase Chain Reaction

The template DNA can be amplified in a *polymerase chain reaction* (PCR) (Fig. 3.6). PCR is based on the interaction of DNA polymerase with DNA. PCR is an iterative process, with each iteration consisting of the following steps: annealing of the short single-stranded DNA molecules, called *primers*, that complementary pair of the templates' ends; extending of the primers in the 5' to 3' direction by DNA polymerase by successively adding nucleotides to the 3' end of the primer; denaturating of the newly elongated double-stranded DNA molecules to separate their strands; and cooling to allow re-annealing of the short, newly amplified single-stranded DNAs. Each cycle

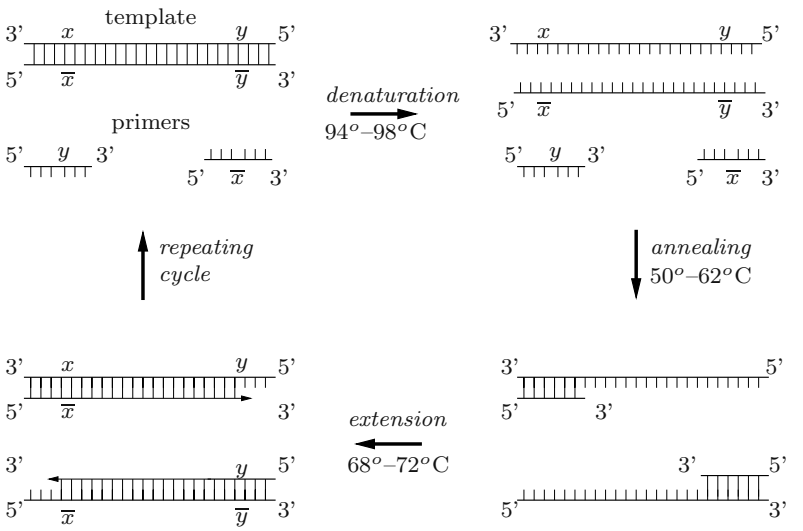

Fig. 3.6 One cycle of PCR.

doubles the number of target DNA molecules. Today, PCR is one of the most fundamental laboratory techniques in modern molecular biology. PCR is based on the interaction of DNA polymerase with DNA.

Parallel overlap assembly (POA) is a method to generate a pool of DNA molecules (combinatorial library). Short single-stranded DNA molecules, called also *oligonucleotides*, overlap after annealing and their sticky ends are extended by DNA polymerase in 5' to 3' direction. Repeated denaturation, annealing, and extension cycles increase the length of the strands. Unlike PCR, where the target DNA strands double in every cycle, in POA, the number of DNA strands does not change, only the length increases with the cycle progression (Fig. 3.7).

The short, single-stranded DNA molecules (oligonucleotides) can be designed by using available software (e.g., DNASequenceGenerator). The GC-contents can be specified as input affecting the melting temperature of the sequences. Oligonucleotides can be synthesized in vitro using PCR.

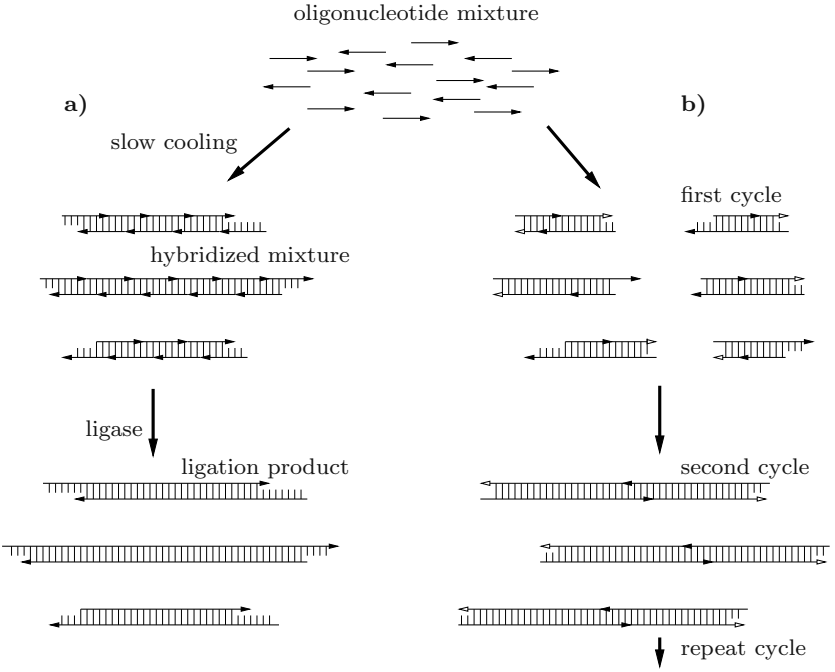

Fig. 3.7 Synthesis methods for combinatorial libraries: **a** Annealing/ligation: The arrow heads indicate the 3' end. **b** POA: The thick arrow represents the synthesized oligomers which are the input of the computation. The thin arrows represent the part that is elongated by DNA polymerase. The arrow heads indicate the 3' ends.

3.2 Physical Chemistry

Computing with biological macromolecules such as DNA is based on a fundamental physicochemical process. Therefore, knowledge about the thermodynamics and kinetics of these processes is necessary.

3.2.1 Thermodynamics

The thermodynamics of physicochemical processes is concerned with energy changes accompanying physical and chemical changes. This section addresses the thermodynamics of DNA pairing and denaturation of DNA molecules.

Nearest Neighbor Model

The relative stability of a double-stranded DNA molecule appears to depend primarily on the identity of the nearest neighbor bases. Ten different nearest neighbor interactions are possible in any double-stranded DNA molecule. These pairwise interactions are `AA/TT`, `AT/TA`, `TA/AT`, `CA/GT`, `GT/CA`, `CT/GA`, `GA/CT`, `CG/GC`, `GC/CG`, and `GG/CC`, denoted in the direction of 5' to 3'/3' to 5'. The relative stability and temperature-dependent behavior of each DNA nearest neighbor interaction can be characterized by Gibbs free energy, enthalpy, and entropy. *Gibbs free energy* describes the potential of a reaction to occur spontaneously; *enthalpy* provides the amount of heat released from or absorbed by the system; and *entropy* measures the randomness or disorder of a system. The corresponding parameters presented in Table 3.1 were derived from J. SantaLucia, Jr. and D. Hicks (2004) in 1 M `NaCl` at temperature 37°C. The Gibbs free energy ΔG° is related to the enthalpy ΔH° and the entropy ΔS° by the standard thermodynamic relationship

$$\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ \; . \tag{3.1}$$

As the Gibbs free energy data listed in Table 3.1 were calculated at 37°C, the ΔG° values at any other temperature can be computed by using the tabulated enthalpy and entropy data. For instance, the relative stability of the `GC/CG` pair at 50°C is $(-9.8\,\text{kcal}) - [(323.15\,\text{K})(-0.0244\,\text{kcal/K})] = -1.915\,\text{kcal}$ per mol compared with $-2.24\,\text{kcal/mol}$ at 37°C.

Gibbs Free Energy

The Gibbs free energy of a double-stranded molecule given by $x = a_1 \ldots a_n$, with reverse complementary strand $\overline{a}_n \ldots \overline{a}_1$, is calculated as

Table 3.1 Nearest neighbor thermodynamics. The units for ΔG° and ΔH° are kcal/mol of interaction, and the unit for ΔS° is cal/K per mol of interaction. The symmetry correction is applied only to self-complementary duplexes. The terminal AT penalty applies to each end of a duplex that has terminal AT. A duplex with both ends closed by AT pairs has a penalty of +1.0 kcal/mol for ΔG°.

Interaction	ΔH°	ΔS°	ΔG°
AA/TT	–7.6	–21.3	–1.00
AT/TA	–7.2	–20.4	–0.88
TA/AT	–7.2	–21.3	–0.58
CA/GT	–8.5	–22.7	–1.45
GT/CA	–8.4	–22.4	–1.44
CT/GA	–7.8	–21.0	–1.28
GA/CT	–8.2	–22.2	–1.30
CG/GC	–10.6	–27.2	–2.17
GC/CG	–9.8	–24.4	–2.24
GG/CC	–8.0	–19.9	–1.84
Initiation	+0.2	–5.7	+1.96
Terminal AT penalty	+2.2	+6.9	+0.05
Symmetry correction	0.0	–1.4	+0.43

$$\Delta G^\circ(x) = \Delta g_i + \Delta g_s + \sum_{i=1}^{n-1} \Delta G^\circ(a_i a_{i+1}/\overline{a}_i \overline{a}_{i+1}) \,, \tag{3.2}$$

where Δg_i denotes the helix-initiation energy and Δg_s is the symmetry correction.

Example 3.1. Consider the double-stranded DNA molecule

$$\begin{aligned} &5' - \texttt{GCAATGGC} - 3' \\ &3' - \texttt{CGTTACCG} - 5' \,. \end{aligned}$$

The Gibbs free energy is given by

$$\begin{aligned} \Delta G^\circ &= \Delta g_i + \Delta g_s + \Delta G^\circ(\texttt{GC/CG}) + \Delta G^\circ(\texttt{CA/GT}) + \Delta G^\circ(\texttt{AA/TT}) \\ &\quad + \Delta G^\circ(\texttt{AT/TA}) + \Delta G^\circ(\texttt{TG/AC}) + \Delta G^\circ(\texttt{GG/CC}) + \Delta G^\circ(\texttt{GC/CG}) \\ &= 1.96 + 0.0 - 2.24 - 1.45 - 1.00 - 0.88 - 1.44 - 1.84 - 2.24 \\ &= -9.13 \text{ kcal/mol.} \end{aligned}$$

♢

The enthalpy of a double-stranded molecule given by $x = a_1 \dots a_n$, with reverse complementary strand $\overline{a}_n \dots \overline{a}_1$, is computed as

$$\Delta H^\circ(x) = \Delta h_i + \sum_{i=1}^{n-1} \Delta H^\circ(a_i a_{i+1}/\overline{a}_i \overline{a}_{i+1}) \,, \tag{3.3}$$

where Δh_i denotes the helix initiation enthalpy. The entropy of a short stretch of double-stranded DNA, also called *duplex*, can either be computed from Table 3.1 or by using Eq. (3.1).

Example 3.2. The double-stranded DNA molecule in the above example has the enthalpy $\Delta H^\circ = -59.2\,\text{kcal/mol}$ and the entropy $\Delta S^\circ = -161.5\,\text{cal/K}$ per mol. Alternatively, the entropy at $T = 37°\text{C}$ can be calculated as

$$\Delta S^\circ = \frac{(\Delta H^\circ - \Delta G^\circ) \cdot 1000}{T} = \frac{(-59.2 + 9.13) \cdot 1000}{310.15} = 161.4\,\text{cal/K per mol.}$$

♢

Melting Temperature

The *melting temperature* is the temperature at which half of the strands in a solution are complementary base-paired and half are not. Melting is the opposite process of hybridization, which is the separation of double strands into single strands. When the reaction temperature increases, an increasing percentage of double strands melt. For oligonucleotides in solution, the melting temperature is given by

$$T_m = \frac{\Delta H^\circ}{\Delta S^\circ + R \ln([C_T]/z)}\,, \tag{3.4}$$

where R is the gas constant, $[C_T]$ is the total molar strand concentration, and z equals 4 for nonself-complementary strands and equals 1 for self-complementary strands. Melting curves can be measured by UV absorbance at 260 nm. With the temperature, the amount of dsDNA decreases (which is paralleled by increase of the amount of ssDNA) and leads to enhancement of the absorbance at 260 nm.

Example 3.3. In view of the above non-self-complementary duplex with strand concentration of 0.2 mM for each strand, the melting temperature is

$$T_m = \frac{-59.2 \cdot 1000}{-161.5 + 1.987 \cdot \ln(0.0004/4)} - 273.15°\text{C} = 56.1°\text{C}\,.$$

♢

3.2.2 Chemical Kinetics

The kinetics of physicochemical processes is concerned with the reaction rates of the reactants. This section addresses specific reactions involving DNA molecules.

Chemical Reactions

A chemical reaction is a process that results in an interconversion of chemical substances. The substances initially involved in a chemical reaction are termed *reactants*. Chemical reactions are usually characterized by a chemical change, and they provide one or more products which are generally different from the reactants.

Consider a spatially homogeneous mixture of m reactants X_i, $1 \leq i \leq n$, which react to provide a mixture of n products Y_j, $1 \leq j \leq l$. This chemical reaction can be formally described by the chemical equation

$$\alpha_1 X_1 + \ldots + \alpha_n X_n \xrightarrow{k} \beta_1 Y_1 + \ldots + \beta_l Y_l \,, \tag{3.5}$$

where α_i and β_j are the stoichiometric coefficients with respect to X_i and Y_j. This reaction states that α_1 molecules of substance X_1 react with α_2 molecules of substance X_2 and so on, to give β_j molecules of substance Y_j, $1 \leq j \leq l$. The reaction (3.5) can be described by the *reaction-rate equation*

$$r = k[X_1]^{\alpha_1} \cdots [X_n]^{\alpha_n}, \tag{3.6}$$

where r is the *reaction rate* (in M/s), k is the *rate constant*, and $[X_i]$ is the concentration (in mol/l) of the reactant X_i. The rate constant $k = k(T)$ is mainly affected by the reaction temperature T as described by the *Arrhenius equation*

$$k = \kappa e^{-E_a/RT} \,, \tag{3.7}$$

where κ is the frequency collision factor, E_a is the activation energy (in kcal/mol) necessary to overcome so that the chemical reaction can take place, and R is the gas constant.

The *order* of a chemical reaction is the power to which its concentration term is raised in the reaction-rate equation. Hence, the order of the reaction (3.5) is given by the term $\alpha = \sum_i \alpha_i$. Generally, reaction orders are determined by experiments. For instance, if the concentration of reactant X_i is doubled and the rate increases by 2^{α_i}, then the order of this reactant is α_i. In view of (3.6), the unit of the reaction constant is $(\mathrm{M/s})/\mathrm{M}^{\alpha}$, where α is the reaction order.

Deterministic Chemical Kinetics

The traditional way of treating chemical reactions in a mathematical manner is to translate them into ordinary differential equations. Suppose that there are sufficient molecules so that the number of molecules can be approximated as a continuously varying quantity that varies deterministically over

time. Then a chemical reaction can be described by a coupled system of differential equations for the concentrations of each substance in terms of the concentrations of all others:

$$\frac{d[X_i]}{dt} = f_i([X_1], \ldots, [X_n]), \quad 1 \le i \le n \; . \tag{3.8}$$

Subject to prescribed initial conditions, these differential reaction-rate equations can only be solved analytically for rather simple chemical systems. Alternatively, these systems can be tackled numerically by using a finite difference method.

Example 3.4. The *Lokta-Volterra system* describes a set of coupled autocatalytic reactions:

$$\begin{aligned} \overline{X} + Y_1 &\stackrel{c_1}{\to} 2Y_1, \\ Y_1 + Y_2 &\stackrel{c_2}{\to} 2Y_2, \\ Y_2 &\stackrel{c_3}{\to} Z \; . \end{aligned}$$

Here the bar over the reactant X signifies that its molecular population level is assumed to remain constant. These reactions also mathematically model a simple predator-prey ecosystem. The first reaction describes how prey species Y_1 reproduces by feeding on foodstuff X; the second reaction explains how predator species Y_2 reproduces by feeding on prey species Y_1; and the last reaction details the eventual demise of predator species Y_2 through natural causes. The corresponding reaction-rate equations are as follows:

$$\begin{aligned} \frac{d[Y_1]}{dt} &= c_1[\overline{X}][Y_1] - c_2[Y_1][Y_2], \\ \frac{d[Y_2]}{dt} &= c_2[Y_1][Y_2] - c_3[Y_2] \; . \end{aligned}$$

♢

Example 3.5. Consider the first-order reaction (e.g., irreversible isomerization or radioactive decay),

$$X \stackrel{k}{\longrightarrow} Y \; .$$

The corresponding differential reaction-rate equation is given by

$$\frac{d[X]}{dt} = -k[X] \; .$$

In view of the initial condition $[X] = X_0$ at $t = 0$, the solution is

$$[X](t) = X_0 e^{-kt} \; .$$

♢

3.2.3 DNA Annealing Kinetics

DNA annealing kinetics describes the reversible chemical reaction of the annealing of complementary single-stranded DNA into double-stranded DNA.

DNA pairing from single-stranded oligonucleotides is described by the chemical equation

$$\mathrm{ssDNA}_1 + \mathrm{ssDNA}_2 \underset{k_r}{\overset{k_f}{\rightleftharpoons}} \mathrm{dsDNA} \ . \tag{3.9}$$

This reaction can proceed in both directions and thus is reversible. The forward (k_f) and reverse (k_r) rate constants describe the forward hybridization reaction and the reverse denaturation reaction, respectively. When the reaction (3.9) reaches the equilibrium, both forward and reverse reaction rates are equal. Then the concentrations are constant and do not change with time.

The forward rate constant k_f depends on DNA length, sequence context, and salt concentration:

$$k_f = \frac{k'_N \sqrt{L_s}}{N} \ , \tag{3.10}$$

where L_s is the length of the shortest strand participating in the duplex formation; N is the total number of base pairs present; and k'_N is the nucleation rate constant, estimated to be $(4.35 \log_{10}[\mathtt{Na}^+] + 3.5) \times 10^5$ where $0.2 \leq [\mathtt{Na}^+] \leq 4.0\,\mathrm{mol/l}$. The reverse rate constant k_r is very sensitive to DNA length and sequence:

$$k_r = k_f e^{\Delta G^\circ / RT} \ , \tag{3.11}$$

where R is the gas constant and T is the incubation temperature. Hybridization in vitro is usually carried out at temperature $T = T_m - 298.15\,\mathrm{K}$, where T_m denotes the melting temperature.

3.2.4 Strand Displacement Kinetics

DNA kinetics has the specific feature that displacement of DNA strands can take place. This is described by the chemical equation

$$A_m/B + B_m \underset{k_r}{\overset{k_f}{\rightleftharpoons}} A_m/B_m + B \ , \tag{3.12}$$

where A_m/B stands for a partially double-stranded DNA molecule; B_m stands for an oligonucleotide; A_m/B_m is the resulting completely

complementary (perfectly paired) double-stranded DNA molecule; and B is the released single DNA strand (Fig. 7.26). This second-order reaction can be described by the differential reaction-rate equation

$$\frac{d[A_m/B_m]}{dt} = \frac{d[B]}{dt} = k_f[A_m/B][B_m] - k_r[A_m/B_m][B] \; . \tag{3.13}$$

The concentration of A_m/B at time t depends on its initial concentration $[A_m/B]_0$ and the concentration of A_m/B_m at time t,

$$[A_m/B] = [A_m/B]_0 - [A_m/B_m] \; . \tag{3.14}$$

Similarly,

$$[B_m] = [B_m]_0 - [A_m/B_m] \; . \tag{3.15}$$

Under appropriate hybridization conditions, the dissociation rate constant k_r of the reverse reaction is neglible. In this way, we obtain

$$\frac{d[A_m/B_m]}{dt} = k_f([A_m/B]_0 - [A_m/B_m])([B_m]_0 - [A_m/B_m]) \; . \tag{3.16}$$

Equivalently, we have

$$\int \frac{d[A_m/B_m]}{([A_m/B]_0 - [A_m/B_m])([B_m]_0 - [A_m/B_m])} = k_f \int dt \; . \tag{3.17}$$

Integration yields the concentration of the product A_m/B_m at time t,

$$[A_m/B_m] = \frac{[A_m/B]_0[B_m]_0(1 - e^{([B_m]_0 - [A_m/B]_0)k_f t})}{[A_m/B]_0 - [B_m]_0 e^{([B_m]_0 - [A_m/B]_0)k_f t}} \; . \tag{3.18}$$

This equation shows that at large time instant t, that is, after the reaction is complete, the concentration of the product A_m/B_m tends towards the concentration of the reactant, either A_m/B or B_m, depending on which of the initial concentrations is lower.

3.2.5 Stochastic Chemical Kinetics

Deterministic chemical kinetics assumes that a chemical reaction system evolves continuously and deterministically over time. But this process is neither continuous, as the molecular population level can change only by a discrete integer amount, nor deterministic, as it is impossible to predict the exact molecular population levels at future time instants without taking into account positions and velocities of the molecules.

Master Equations

In view of the shortcomings of deterministic chemical kinetics, the time evolution of a chemical system can be alternatively analyzed by a kind of random-walk process. This process can be described by a single differential-difference equation known as a master equation.

Suppose that there is a container of volume V containing a spatially uniform mixture of n chemical substances which can interact through m specific chemical reactions. This chemical system can be represented by the probability density function $P(X_1, \ldots, X_n; t)$, which denotes the probability that there will be X_i molecules of the ith substance in volume V at time t, $1 \leq i \leq n$. The knowledge of this function would provide a complete stochastic characterization of the system at time t. In particular, the *kth moment* of the probability density function P with respect to X_i, $1 \leq i \leq n$, is given by

$$X_i^{(k)}(t) = \sum_{X_1=0}^{\infty} \cdots \sum_{X_n=0}^{\infty} X_i^k P(X_1, \ldots, X_n; t), \quad k \geq 0 \,. \tag{3.19}$$

The first and second moments are of special interest. While the *mean* $X_i^{(1)}(t)$ provides the average number of molecules of the ith substance in volume V at time t, the *root-mean-square deviation* that occurs about this average is given by

$$\Delta_i(t) = \sqrt{X_i^{(2)}(t) - [X_i^{(1)}(t)]^2} \,. \tag{3.20}$$

In other words, we may expect to find between $X_i^{(1)}(t) - \Delta_i(t)$ and $X_i^{(1)}(t) + \Delta_i(t)$ molecules of the ith substance in volume V at time t.

The master equation describes the time evolution of the probability density function $P(X_1, \ldots, X_n; t)$. For this, let $a_\mu dt$ denote the probability that an R_μ reaction will occur in volume V during the next time interval of length dt given that the system is in state $(X_1, \ldots, X_n)$ at time t, $1 \leq \mu \leq m$. Moreover, let $b_\mu dt$ denote the probability that the system undergoes an R_μ reaction in volume V during the next time interval of length dt, $1 \leq \mu \leq m$. Then the time evolution of the chemical system can be described by the *master equation*

$$P(X_1, \ldots, X_n; t + dt) = P(X_1, \ldots, X_n; t)[1 - \sum_{\mu=1}^{m} a_\mu dt] + \sum_{\mu=1}^{m} b_\mu dt. \tag{3.21}$$

The first term is the probability that the system will be in the state $(X_1, \ldots, X_n)$ at time t and will remain in this state during the next time interval of length dt. The second term provides the probability that the

system undergoes at least one R_μ reaction during the next time interval of length dt, $1 \leq \mu \leq m$. The master equation can be equivalently written as

$$\frac{\delta}{\delta t}P(X_1,\ldots,X_n;t) = \sum_{\mu=1}^{m}[b_\mu - a_\mu P(X_1,\ldots,X_n;t)] \,. \tag{3.22}$$

The probability density $a_\mu dt$ can be expressed by another probability density. For this, let h_μ be the random variable which specifies the number of distinct molecular reactant combinations for the reaction R_μ present in volume V at time t, $1 \leq \mu \leq m$ (Table 3.2). Moreover, let c_μ be the so-called *stochastic reaction constant* depending only on the physical properties of the molecules and the temperature of the system, so that $c_\mu dt$ is the average probability that a particular combination of R_μ reactant molecules will react in the next time interval of length dt, $1 \leq \mu \leq m$. Thus,

$$a_\mu dt = h_\mu c_\mu dt, \quad 1 \leq \mu \leq m \,. \tag{3.23}$$

The stochastic reaction constants depend on the type of chemical reaction. To this end, notice that a chemical reactant X has $x = N_A[X]V$ molecules in a volume of V litres, where N_A is the Avogadro number. For instance, the first-order reaction $X \xrightarrow{k} Y$ amounts to the reaction-rate equation $r = k[X]\,\mathrm{M/s}$. The reaction decreases X at a rate of $N_A k[X]V = kx$ molecules per second and delivers cx molecules per second. Thus, $c = k$. The second-order reaction $X + Y \xrightarrow{k} Z$ gives rise to the reaction-rate equation $r = k[X][Y]\,\mathrm{M/s}$. The reaction proceeds at a rate of $N_A k[X][Y]V = kxy/(N_A V)$ molecules per second and provides cxy molecules per second. Thus, $c = k/(N_A V)$.

Example 3.6. Reconsider the Lokta-Volterra system in Example 3.4. The corresponding master equation is given by

$$\begin{aligned}\frac{\delta}{\delta t}P = {} & c_1[(Y_1-1)P(X,Y_1-1,Y_2,Z;t) - Y_1 P] \\ & + c_2[(Y_1+1)(Y_2-1)P(X,Y_1+1,Y_2-1,Z;t) - Y_1 Y_2 P] \\ & + c_3[(Y_2+1)P(X,Y_1,Y_2+1,Z-1;t) - Y_2 P] \,,\end{aligned}$$

Table 3.2 Reaction R_μ and corresponding random variable h_μ.

Reaction R_μ	Random Variable h_μ
$\emptyset \rightarrow$ products	$h_\mu = 1$
$A_1 \rightarrow$ products	$h_\mu = X_1$
$A_1 + A_2 \rightarrow$ products	$h_\mu = X_1 X_2$
$2A_1 \rightarrow$ products	$h_\mu = X_1(X_1-1)/2$
$A_1 + A_2 + A_3 \rightarrow$ products	$h_\mu = X_1 X_2 X_3$
$A_1 + 2A_2 \rightarrow$ products	$h_\mu = X_1 X_2 (X_2-1)/2$
$3A_1 \rightarrow$ products	$h_\mu = X_1(X_1-1)(X_1-2)/6$

where $P = P(X, Y_1, Y_2, Z; t)$ is assumed to be independent of reactant X. $\diamondsuit$

Example 3.7. Consider the first-order reaction

$$X \xrightarrow{c} Y \ .$$

The corresponding master equation has the form

$$\frac{\delta}{\delta t} P(X;t) = c[(1 - \delta_{X_0,X})(X+1)P(X+1;t) - XP(X;t)] \ .$$

In view of the initial condition $P(X;0) = \delta_{X,X_0}$, the master equation can be solved analytically yielding the standard binomial probability function

$$P(X;t) = \binom{X_0}{X} e^{-cXt}[1 - e^{-ct}]^{(X_0 - X)}, \quad 0 \le X \le X_0 \ .$$

The mean of the probability function is given by

$$X^{(1)}(t) = X_0 e^{-ct}$$

and the root-mean-square deviation turns out to be

$$\Delta(t) = \sqrt{X_0 e^{-ct}(1 - e^{-ct})} \ .$$

$\diamondsuit$

The master equation is mathematically tractable only for simple chemical systems. Fortunately, there is a way to evaluate the time behavior of a chemical system without having to deal with the master equation directly.

Gillespie's Direct Reaction Method

Suppose the chemical system is in state $(X_1, \ldots, X_n)$ at time t. In order to drive the system forward, two questions must be answered: When will the next reaction occur? What kind of reaction will occur?

To this end, consider the so-called *reaction probability density function* $P(\tau, \mu)$ so that $P(\tau, \mu)d\tau$ is the probability that given the state $(X_1, \ldots, X_n)$ at time t, the next reaction will occur in the interval $(t + \tau, t + \tau + d\tau)$, and will be an R_μ reaction. Observe that $P(\tau, \mu)$ is a joint probability density function of continuous variable τ, $\tau \ge 0$, and discrete variable μ, $1 \le \mu \le m$.

Theorem 3.8. *If* $a_0 = \sum_\nu a_\nu$, *then*

$$P(\tau, \mu) = a_\mu \exp\{-a_0 \tau\}, \quad 0 \le \tau < \infty, \ 1 \le \mu \le m \ . \tag{3.24}$$

Proof. Let $P_0(\tau)$ be the probability that given the state $(X_1, \ldots, X_n)$ at time t, no reaction will occur during the next interval of length τ. Then we have

$$P(\tau, \mu)d\tau = P_0(\tau) \cdot a_\mu d\tau \ . \tag{3.25}$$

But $1 - \sum_\nu a_\nu d\tau'$ is the probability that in state $(X_1, \ldots, X_n)$, no reaction will occur during the next time interval of length $d\tau'$. Thus

$$P_0(\tau' + d\tau') = P_0(\tau') \cdot [1 - \sum_\nu a_\nu d\tau'] \tag{3.26}$$

and hence

$$P_0(\tau) = \exp\{-\sum_\nu a_\nu \tau\} \ . \tag{3.27}$$

Substituting this expression for $P_0(\tau)$ into Eq. (3.25) yields the result. □

The direct reaction method is based on the decomposition of the reaction probability density function $P(\mu, \tau)$. This technique is termed conditioning and leads to the equation

$$P(\tau, \mu) = P_1(\tau)P_2(\mu \mid \tau) \ . \tag{3.28}$$

Here $P_1(\tau)d\tau$ is the probability that the next reaction will occur in the interval $(t + \tau, t + \tau + d\tau)$, and $P_2(\mu \mid \tau)$ is the probability that the next reaction will be an R_μ reaction, given that the next reaction will occur at time $t + \tau$.

The probability $P_1(\tau)d\tau$ is the sum of the probabilities $P(\tau, \mu)d\tau$ over all μ-values. Thus, in view of Eq. (3.24),

$$P_1(\tau) = a_0 \exp\{-a_0\tau\}, \quad 0 \le \tau < \infty \ . \tag{3.29}$$

Substituting this into Eq. (3.28) yields the discrete probability function

$$P_2(\mu \mid \tau) = \frac{a_\mu}{a_0}, \quad 1 \le \mu \le m \ . \tag{3.30}$$

As $P_2(\mu \mid \tau)$ is independent of τ, we put $P_2(\mu) = P_2(\mu \mid \tau)$.

The direct reaction algorithm belongs to the class of Monte Carlo methods and simulates the stochastic process described by the probability density function $P(\tau, \mu)$. For this, the stochastic process proceeds in discrete time steps choosing in the actual time instant t a pair (τ, μ) according to density $P(\tau, \mu)$ so that the reaction R_μ occurs at time instant $t + \tau$, and $t + \tau$ becomes the actual time instant. But in view of Eq. (3.28) and the remark in the last paragraph, $P(\tau, \mu)$ can be viewed as a joint probability density function

$$P(\mu, \tau) = P_1(\tau)P_2(\mu) \ . \tag{3.31}$$

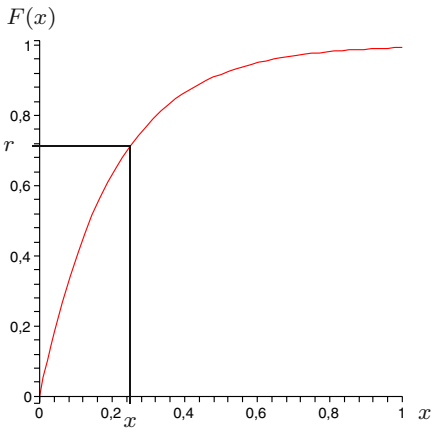

Fig. 3.8 Inversion method.

Thus, a pair (τ, μ) can be drawn so that τ is chosen according to probability density $P_1(\tau)$ and μ is taken according to probability function $P_2(\mu)$. To this end, the *inversion method* (Fig. 3.8) is employed:

> In order to generate a random value x according to a given probability density function $P(x)$, draw a random number r from the uniform distribution in the unit interval $[0, 1]$ so that $F(x) = r$, where $F(x) = \int_{-\infty}^{x} P(u)du$ is the corresponding probability distribution function.

First, the distribution function of the density $P_1(\tau)$ is given by

$$F_1(\tau) = \int_{-\infty}^{\tau} P_1(\tau')d\tau' = 1 - \exp\{-a_0\tau\} \; . \tag{3.32}$$

Take a random number r_1 from the uniformly distributed unit interval and put $F_1(\tau) = r_1$. Resolving for τ (and replacing the random variable $1 - r_1$ by the statistically equivalent random variable r_1) yields

$$\tau = (1/a_0)\log(1/r_1) \; . \tag{3.33}$$

Second, the (discrete) distribution function of the probability function $P_2(\mu)$ is given as follows:

$$F_2(\mu) = \sum_{\nu=-\infty}^{\mu} P_2(\nu) = \sum_{\nu=1}^{\mu} a_\nu/a_0 \; . \tag{3.34}$$

Draw a random number r_2 from the uniformly distributed unit interval and take for μ that value which satisfies

$$F_2(\mu - 1) < r_2 \leq F_2(\mu) \,. \tag{3.35}$$

That means, take μ to be that integer for which

$$\sum_{\nu=1}^{\mu-1} a_\nu < r_2 a_0 \leq \sum_{\nu=1}^{\mu} a_\nu \,. \tag{3.36}$$

These observations lead to Gillespie's direct reaction algorithm 3.1 (1977).

Example 3.9. Consider the DNA hybridization reaction

$$\mathrm{ssDNA}_1 + \mathrm{ssDNA}_2 \xrightarrow{k} \mathrm{dsDNA} \,.$$

Algorithm 3.1 Gillespie's Direct Reaction Method

Input: Stochastic reaction constants $c_1 \ldots, c_m$, initial molecular population numbers $X_1, \ldots, X_n$

1: $t \leftarrow 0$
2: **for** $i \leftarrow 1$ to N **do**
3: Calculate the propensities $a_\nu = h_\nu c_\nu$, $1 \leq \nu \leq m$.
4: Generate random numbers r_1 and r_2 from the uniformly distributed unit interval.
5: Calculate τ and μ according to (3.33) and (3.36), respectively.
6: $t \leftarrow t + \tau$
7: Adjust the molecular population levels to reflect the R_μ reaction.
8: **end for**

Take 100,000 oligonucleotides of length 23 nt in a volume V of 10^{-15} l. This gives an approximate concentration of 1.66×10^{-4} mol/l. The melting temperature T_m was set to 338.15 K and the $[\mathtt{Na}^+]$ value was taken to be 4.0 mol/l. The simulation of the reaction by Gillespie's direct reaction method shows that the formation of double-stranded DNA is favored (Fig. 3.9). ♢

Gillespie's First Reaction Method

Gillespie's algorithm is direct in the sense that it calculates the quantities τ and μ directly. D. Gillespie developed another simulation algorithm which generates a putative time τ_μ for each reaction R_μ to occur and lets μ be the reaction whose putative time comes first and lets τ be the putative time τ_μ. To this end, the putative times τ_μ are drawn according to Eq. (3.33),

$$\tau_\mu = (1/a_0) \log(1/r_\mu) \,, \tag{3.37}$$

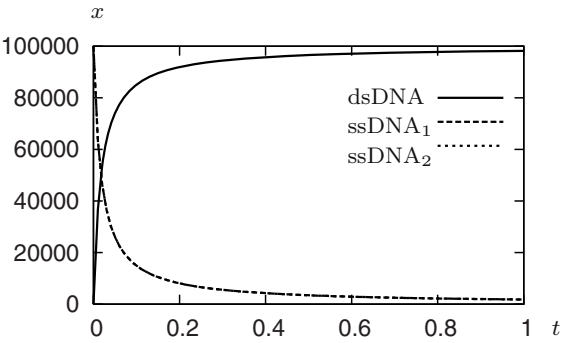

Fig. 3.9 Results of stochastic simulation of DNA hybridization reaction (time (h) vs. number of molecules).

where r_μ is a random number from the uniformly distributed unit interval, $1 \leq \mu \leq m$. These observations lead to Gillespie's first reaction algorithm 3.2 (1976). Both algorithms are equivalent in the sense that the probability distributions used to choose the pair (μ, τ) are the same. A more efficient version of Gillespie's first reaction algorithm is the Gibson-Bruck algorithm (2000).

Algorithm 3.2 Gillespie's First Reaction Method

Input: Stochastic reaction constants $c_1 \ldots, c_m$, initial molecular population numbers $X_1, \ldots, X_n$

1: $t \leftarrow 0$
2: **for** $i \leftarrow 1$ to N **do**
3: Calculate the propensities $a_\nu = h_\nu c_\nu$, $1 \leq \nu \leq m$.
4: **for** $\nu \leftarrow 1$ to m **do**
5: Generate putative time τ_ν according to (3.37).
6: **end for**
7: Let R_μ be the reaction whose putative time τ_μ is smallest.
8: $t \leftarrow t + \tau_\mu$
9: Adjust the molecular population levels to reflect the R_μ reaction.
10: **end for**

3.3 Genes

The gene is considered to be the basic unit of inheritance. The genetic material carries the information that directs all physiological activities of the cell and specifies the developmental changes of the multicellular organisms. Understanding the gene structure and function is therefore of fundamental importance.

3.3.1 Structure and Biosynthesis

The stored information in the DNA sequence is expressed in a two-stage process, comprising *transcription* into RNA and *translation* of the nucleotide sequence into an amino acid sequence. The RNA is another type of nucleic acid that differs from the DNA in its sugar (ribose instead of deoxyribose), and the thymine (`T`) base is replaced by uracil (`U`). RNA is usually single-stranded and can have different functions in the cell, for example, transfer RNA (tRNA, transfers amino acid to the polypeptide chain) or ribosomal RNA (rRNA, one of the components of the ribosomes). The RNA that transfers the information from the DNA to the ribosomes, which represent the protein synthesis machinery of the cell, is called *messenger RNA* (mRNA). One of the DNA strands, the *template strand*, directs the synthesis of the mRNA via complementary base-pairing by the addition of nucleotides to the 3' end to the growing mRNA (5' to 3' direction of synthesis). The process is controlled by some functional sequence regions that serve as a set of rules to govern the transcription (Fig. 3.10). The transcription is initiated from a promoter located upstream from the DNA coding sequence. Promoter sequences are highly conserved and are recognized by transcription factors, which recruit further the *RNA polymerase*, the enzyme that is responsible for transcribing the coding part of the DNA into mRNA. Different proteins that participate in the transcription process bind to a specific sequence in the double-stranded DNA and usually make contacts with the bases in the major groove of the DNA helix. During elongation, the RNA polymerase moves along the DNA and elongates the RNA, whereas sequential unwinding of the DNA helix precedes the transcription into RNA. Behind the RNA polymerase, the unwound regions base-pair again and restore the original DNA double helix. The process is terminated by a *terminator sequence* located downstream from the coding sequence. Transcription is a fast process: 20 to 50

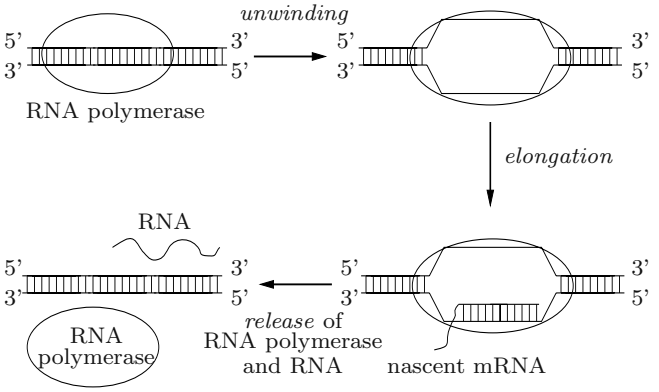

Fig. 3.10 Transcription of DNA into mRNA by RNA polymerase.

nucleotides per second are added to the growing nascent mRNA chain in vivo at 37°C.

In many cases, the function of the single stranded RNA requires a well-defined three-dimensional structure, which also involves base-pairing. In the double-stranded DNA molecules, each base is fixed to its corresponding partner to form a complementary base pair. For the single-stranded RNA (valid also for the single-stranded DNA), each base can bind only one partner at a time, but there are multiple potential partners for such pairing within one chain (Fig. 3.11). With different pairing partners along the chain, various intramolecular, partially double-stranded structures can be formed; each of them is stabilized by the effective free energy of the double-stranded complementary region and the type and the length of the loop enclosed in this structure. The complementary structures formed are imperfect and the integrity can be interrupted by several non-complementary regions, forming three types of loops:

- Hairpin loops: A single chain flips back through a non-complementary region and forms a double strand with adjacent sequences.
- Internal loops: Short regions within a long double-stranded region are not complementary.
- Bulge loops: One of the strands contains bases which are not complementary to the opposite sequence.

The formation of complementary stretches releases free energy (negative value of the Gibbs free energy), which accounts for stabilization, whereas the loops introduce a positive value of free energy (i.e., require energy to be formed). Additional energy is released through the hydrophobic interactions between the base pairs, which are stacked over each other within the double-stranded region. The overall stability of the secondary structure of the single-stranded polynucleotide molecule is determined by the sum of the stabilization through base-pairing and destabilization by the loop structures. The free energy has to reach a sufficient negative value overall, or the secondary structure will be not formed. Intramolecular base-pairing is implicated in the termination of the transcription: a hairpin of palindromic sequences at the 3' terminus causes RNA polymerase to pause and terminate the transcription.

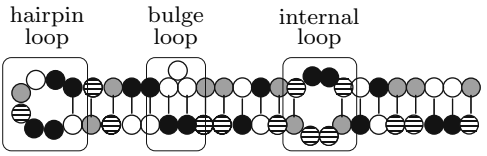

Fig. 3.11 Various intermolecular loops formed by intramolecular base-pairing within a single-stranded RNA molecule. Symbols: open circle = `A`, black circle = `U`, gray circle = `C`, and shaded circle = `G`.

A *gene* is determined by a segment of the DNA comprising the information necessary to be transcribed into functional RNA and further translated into an amino acid sequence with the flanking regulatory and controlling elements that ensure the fidelity of these processes. The information in the mRNA is read from the 5' to 3' direction in groups of three nucleotides and each trinucleotide or triplet is called a *codon*. The starting point of the translation determines the sequence of the non-overlapping codons, which provides the *reading frame*. A mutation that changes the triplet frame by insertion or deletion of base pairs will cause a change in the reading frame, known as *frameshift*. The new reading frame will generate a completely new RNA sequence beyond the site of mutation. Adverse environmental conditions or errors during replication are sources of mutations. Mutations are rare stochastic events and any base pair can be mutated. A change of only a single base pair is called *point mutation*. The average spontaneous mutation rate corresponds to changes at individual nucleotides of 10^{-9} to 10^{-10} per generation. Substitution mutations without any apparent effect on the amino acid level are designated as *silent mutations*.

In prokaryotes and in some nematodes the same regulatory sequences can govern the transcription of more than one structural gene (i.e., translation into more than one protein). This functional unit of the DNA sequence that comprises many structural genes controlled from a common promoter and transcribed into one mRNA is called *operon*. The activity of the structural genes within an operon is regulated by an operator, a sequence located downstream from the promoter and upstream from the initial `AUG` codon that determines the start of the mRNA. The operator interacts with the repressor and activator proteins which regulate the transcription of the structural genes and can be encoded within the operon or elsewhere in the genome.

While the prokaryotic genes are colinear with the proteins (i.e., the DNA sequence corresponds exactly to the amino acid sequence), the coding sequence of eukaryotic genes is interrupted by additional non-coding sequences. The coding segments are called *exons* and the non-coding *introns*. The introns are excised from the mRNA and the exons are joined through a process known as *gene splicing* that occurs in the nucleus of the eukaryotic cell. Some mRNAs undergo a self-splicing, but the splicing of the majority of the mRNA is catalyzed by the *spliceosome*, a large RNA-protein complex composed of small nuclear ribonucleo-proteins. Specific recognition signals are located within the intron (e.g., an invariant `GU` at the intron's 5' boundary and invariant `AG` at its 3' boundary define the splice junction). Within the introns, around 20 to 50 residues upstream from the 3' splice site is found a conserved sequence in all the vertebrate mRNAs: `CURAY`, where `R` represents purines (`A` or `G`) and `Y` represents pyrimidines (`C` or `U`). This sequence represents the *branch point*. The splicing is initiated by release of the 5' intron end and joining it to the branch point thereby forming a loop structure, a lariat intermediate. On the second stage, the 3' hydroxyl group of the exon performs a nucleophilic attack at the 3' splice site and ligation of the exons.

The splicing process is not uniform and the exons may be arranged in an alternative pattern, which is known as *alternative splicing*. This process occurs in the higher eukaryotes and is an important mechanism in tissue differentation and developmental regulation of gene expression. Selection of different splicing sites is regulated by serine/arginine residue proteins or SR proteins. Differential use of promoters or termination sites can produce alternative N-termini or C-termini, respectively, in proteins (Fig. 3.12). Alternative reshuffling of the exons or retaining of some of the introns leads to variations in the codons or to a new amino acid sequence. Alternative splicing is an efficient way to economically store larger amounts of genetic information on shorter DNA sequences.

3.3.2 DNA Recombination

DNA recombination refers to a process by which a DNA segment from one DNA molecule is exchanged with a DNA segment from another. A common type of recombination is homologous recombination, known also as *DNA crossover* (Fig. 3.13). When two homologous parts of the DNA align side by

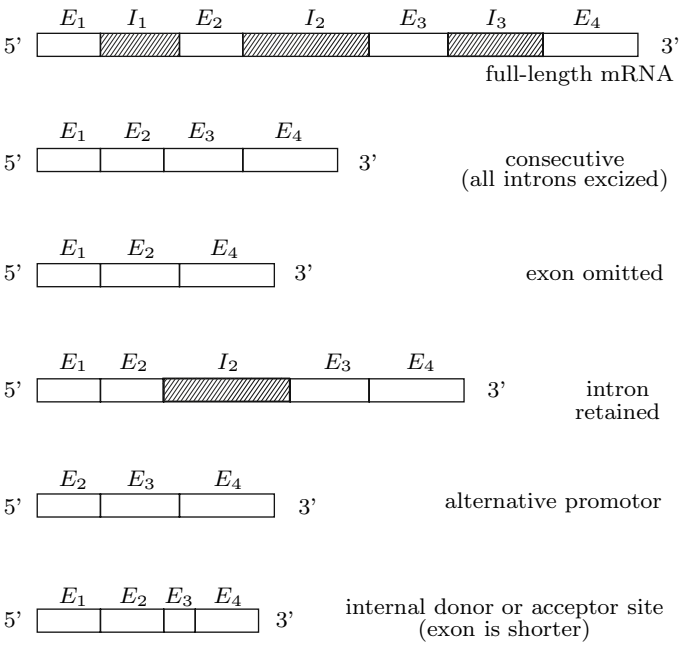

Fig. 3.12 Splicing of mRNA. Introns (I_i) are shaded and exons (E_i) are presented as open structures.

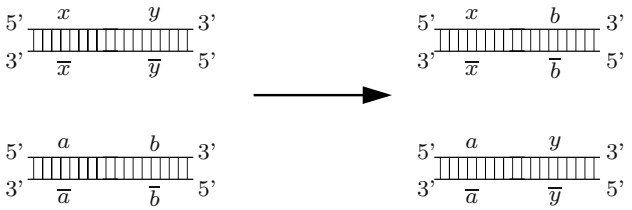

Fig. 3.13 DNA crossover.

side, they can exchange identical parts. In the regions with largely repetitive segments, the alignment might be only partial, which will lead to an unequal crossover. This type of recombination involves the exchange between two homologous DNA molecules, without altering the overall genetic information. The exchange of the strands between homologous DNAs that form heteroduplexes is promoted and guided by specific cellular enzymes. In contrast to the general homologous recombination which necessitates extensive stretches of sequence homology, a site-specific recombination can occur between two sequences with only a small core of homology. The proteins that mediate this process recognize specific target sequences within the DNA, unlike the complementary base-pairing by the homologous recombination. The site-specific recombination creates diversity, thus increasing the array of the proteins that can be synthesized from a certain pool of DNA information. In addition to increasing the genetic diversity, recombination in general plays an essential role for repairing damaged DNA.

3.3.3 Genomes

The complete information carried by the DNA (coding and non-coding sequences) in one organism is referred to as *genome*. In the eukaryotic systems the term genome is specifically applied to the DNA encoded in the nucleus, but the term can also be used for some organelles that contain their own DNA, such as mitochondrial or chloroplast genomes. The total amount of the DNA in one genome is a characteristic feature of each organism and is known as *C-value*, which is defined as the length of the genome. The genomes vary from 10^3 base pairs for some DNA viruses to 10^{11} base pairs for some plants and amphibians. The size of the human genome, for example, is 3×10^9 bp. Large variations in the C-value between similar species are observed: For the amphibious species, the smallest genome is 10^3 bp while the largest is 10^{11} bp. The genome size increases with the rise in complexity of the organisms from prokaryotes to eukaryotes, although in the higher eukaryotes, the proportional correlation between genome size and organism complexity disappears.

The prokaryotic *Escherichia coli* genome consists of 4.7 million base pairs and codes for about 3,000 genes, whereas the 2.9 billion base-pairs haploid human genome is estimated to code for approximately 40,000 genes. The genomes of some lungfishes, however, are larger than those of mammals. Variations in the genome size and in the C-value do not bear the complexity of the organism which gave rise to the C-value paradox. The puzzle of the disproportional C-value to the organism's complexity has been partially solved after identification of non-coding DNA and repetitive DNA sequences.

3.4 Gene Expression

Proteins are the main active players in the biochemical and physiological processes of the cell and implement the unique information that is stored in the ribonucleotid sequences. This task is executed with high fidelity and many steps of control assure the high accuracy of the process.

3.4.1 Protein Biosynthesis

The information in the mRNA is processed in a sequential manner in the 5' to 3' direction, where subsequent codons are translated into amino acids. Each set of three nucleotides corresponds to a specific amino acid, and this genetic code is nearly universal for all living organisms. The four nucleotides (`A`, `U`, `C`, and `G`) at each of the three positions of a codon form 64 possible codons that encode for only 20 standard amino acids and three non-transcribed *non-sense* or *stop-codons* (`UAG`, `UAA`, `UGA`). Hence, the genetic code is redundant and highly degenerate, and multiple codons encode the same amino acid. Six codons exist for the amino acids arginine (`CGU`, `CGC`, `CGA`, `CGG`, `AGA`, `AGG`), leucine (`UUA`, `UUG`, `CUU`, `CUC`, `CUA`, `CUG`), and serine (`AGU`, `AGC`, `UCU`, `UCC`, `UCA`, `UCG`), and the rest of the amino acids are specified by either four, three or two codons. Only two of the amino acids, methionine and tryptophan, are encoded by a single codon (Table 3.3). The set of codons coding for one amino acid differ mostly in the third position (i.e., the mutation at the third position is phenotypically silent), and the degeneracy of the genetic code might be an evolutionarily acquired tolerance to minimize the deleterious effect of point mutations.

The genetic code is widespread but not universal. The genetic code of mitochondria shows some deviations from the "standard" genetic code for some amino acids. In certain proteins, substituted amino acids can be integrated via standard stop codons: `UGA` can code for selenocysteine and `UAG` can code for pyrolysine. The standard genetic code has been expanded to allow incorporation of unnatural amino acids, including amino acids modified with fluorophores for specific detection and chemical and photo-chemical reactive

Table 3.3 The genetic code.

1-Letter Code	3-Letter Code	Codon	1-Letter Code	3-Letter Code	Codon
A	Ala	GCA	L	Leu	CUC
		GCC			CUG
		GCG			CUU
		GCU	K	Lys	AAA
R	Arg	AGA			AAG
		AGG	M	Met	AUG
		CGA	F	Phe	UUC
		CGC			UUU
		CGG	P	Pro	CCA
		CGU			CCC
N	Asn	AAC			CCG
		AAU			CCU
D	Asp	GAC	S	Ser	AGC
		GAU			AGU
C	Cys	UGC			UCA
		UGU			UCC
E	Glu	GAA			UCG
		GAG			UCU
Q	Gln	CAA	T	Thr	ACA
		CAG			ACC
G	Gly	GGA			ACG
		GGC			ACU
		GGG	W	Trp	UGG
		GGU	Y	Tyr	UAC
H	His	CAC			UAU
		CAU	V	Val	GUA
I	IIe	AUA			GUC
		AUC			GUG
		AUU			GUU
L	Leu	UUA	Stop	Ochre	UAA
		UUG		Amber	UAG
		CUA		Opal	UGA

groups for crosslinking. These amino acids are site-specifically incorporated into the peptide chain in response to an amber (UAG) stop codon by an amber suppressor tRNA that is aminoacylated with the desired unnatural amino acid. Another approach of introducing unnatural amino acid relies on the read-through of four-base codons via an aminoacyl-tRNA with an engineered anti-codon loop to accommodate four bases.

Translation of the mRNA is carried out by the ribosomes and each codon binds in a complementary manner to the unpaired bases, known as *anti-codon* of the corresponding tRNA (Fig. 3.14). Transfer RNA is a small non-coding RNA chain (60 to 95 nucleotides) with a cloverleaf secondary structure. It is the carrier for the amino acid, which after the base pairing of the codon and anti-codon is covalently linked to the carboxyl terminus of the growing

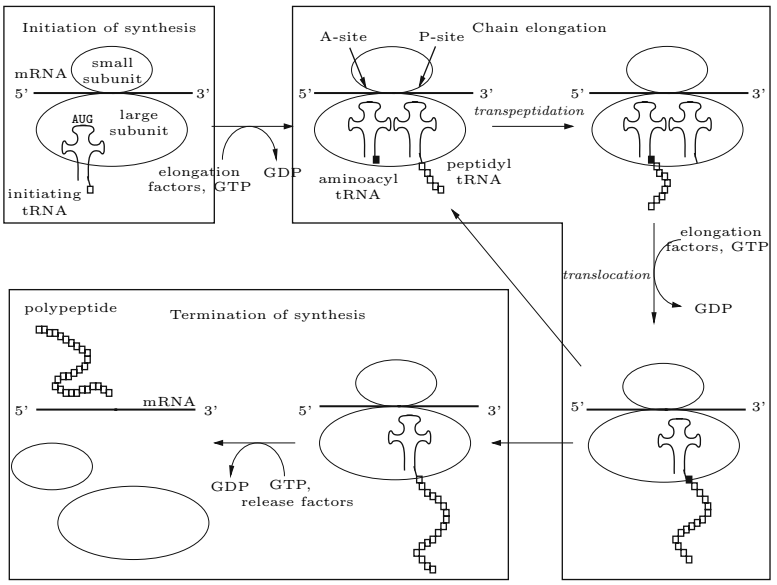

Fig. 3.14 Translation of the genetic information from mRNA to polypeptide chain.

polypeptide chain. The proper tRNA is selected based on the complementary codon-anticodon interactions. Although many organisms contain different isoaccepting tRNAs (different tRNAs specific for the same amino acid), in some cases the same tRNA can bind to two or three codons, encoding the same amino acid. The codon-anticodon pairing in this case is incomplete with a non-Watson-Crick geometry at the third position, known as a *wobble base pair*. Only certain wobble base-pairs are allowed (`G-U` and `I-U`, `I-A`, `I-C`, where `I` is a modified base) whose thermodynamic stability is similar to the Watson-Crick base pair.

Ribosomes are large nucleoprotein complexes consisting of two subunits that read the genetic code on the mRNA in the 5' to 3' direction and translate it into amino acids. Initiation of the translation is a complex process, which requires initiation factors that help the small and large subunits of each ribosome to assemble on the mRNA with the first aminoacyl-tRNA (Fig. 3.14). This process is slow and rate-limiting for the translation. The translation starts at the `AUG` codon, which is recognized by the initiator tRNA for methionine. In eukaryotic cells only initiator tRNA can bind to the small subunit of ribosomes before it assembles with the large subunit on the mRNA. In prokaryotic organisms the recognition of the first `AUG`-codon is controlled by a specific ribosome-binding nucleotide sequence (6 bp) upstream from the start `AUG`-codon. Therefore, in bacteria the translation can be initiated at any `AUG`-position in the mRNA given the presence of the upstream ribosome-binding sequence. As a consequence, the mRNA in

prokaryotes is often polycystronic (i.e., one mRNA molecule can encode for many proteins). Ribosomes elongate the polypeptide chain by adding one amino acid residue at the time in a three-stage reaction cycle (Fig. 3.14). The tRNA charged with an amino acid binds to the acceptor site (A site) of the ribosome at the $n+1$-th codon. The part of the tRNA that carries the amino acid is in the large subunit, whereas the anti-codon at the other end binds to the mRNA codon. The n-th codon that has been read is in the donor site (P site), which is occupied by the peptidyl-tRNA carrying the nascent amino acid chain. A peptide bond is formed in the second stage through a nucleophilic displacement of the peptidyl-tRNA by the 3'-linked amino acid of the aminoacyl-tRNA. The reaction occurs in ATP-independent manner and the energy is provided by the high energy bond between the polypeptide and the peptidyl-tRNA. The new amino acid from the tRNA is transferred to the C-terminus at the growing chain and the *transpeptidation* is catalyzed by the peptidyltranferase activity of the large subunit. In the third and final stage, the new peptidyl-tRNA in the A site is transferred, together with the bound codon of mRNA, to the P site. The efficiency of this translocation process is maintained by an elongation factor that binds to the ribosome together with GTP, delivering the energy for the transfer reaction. The elongation of the polypeptide nascent chain is the most rapid step in the protein synthesis. When the ribosomes encounter the stop codon the synthesis is terminated, which results in a release of the polypeptide and dissociation of the two ribosomal subunits from the mRNA. The termination is facilitated by release factors and GTP-hydrolysis. In the whole biosynthetic cycle of the polypeptide chain, GTP acts as an energy donor, ensuring fastness and irreversibility of the coupled initiation, elongation and termination of the translation.

3.4.2 Proteins – Molecular Structure

Proteins are linear polymers which have to acquire a correct 3D structure to accomplish the physiological functions in the cell. The ribosomes decode the information from the mRNA and translate it into a polypeptide chain, built from the 20 amino acids. The linear sequence of the amino acids that is determined from the genetic information stored in the DNA determines the *primary protein sequence*. The amino acids are linked in a dehydration reaction (release of water) by the C-terminal COOH group of the n-th amino acid and the N-terminal NH_2 group of the $n+1$-th amino acid forming a covalent peptide bond ($-CO-NH-$). The backbone of each protein is identical, while the side chains of the different amino acids introduce diversity into the physicochemical properties of each protein. The differences in the chemical structure of the individual amino acids determines the directionality of the protein chain; the peptide chain starts with a free amino group (N-terminus)

of the first amino acid and ends with a free carboxyl group of the last amino acid (C-terminus).

Short or large sequence regions are forming secondary structures stabilized by non-covalent interactions (i.e., hydrogen bonds, hydrophobic interactions) (Fig. 3.15). The most common structures are helices and beta-sheets. Helices are stabilized with hydrogen bonds between the C = O group of the peptide bond of the n-th residue with the NH group of the $n + 4$-th amino acid. In addition, the tight packing of the helix allows van der Waals interactions of the atoms across the helix. The most common α-helix is right handed and can accommodate on average 3.6 residues per turn. Helices in the proteins range from four to over forty residues, but a typical helix spans about 12 residues, corresponding to over three helical turns.

The second structural element is the beta pleated sheet consisting of strands connected with a hydrogen bond network between NH groups in the backbone of one strand and C = O groups of the adjacent strand, forming a twisted pleated sheet. The backbone hydrogen bonding of the beta sheets is generally considered as slightly stronger than that found in the α-helices. The two neighboring chains can run in the same (parallel beta pleated sheet) or in the opposite direction (anti-parallel beta pleated sheet). Helices and beta-sheets comprise around half of the local structures in the globular proteins. The remaining parts have either coil or loop conformation. Turns and loops establish the joints between different secondary structural elements and almost always occur at the surface of the protein. Turns are also stabilized by a hydrogen bond, either: between the n-th and $n + 3$-th residues (β-turn); between the n-th and $n + 2$-th residues (γ-turn); between the n-th and $n + 4$-th residues (α-turn); between the n-th and $n + 5$-th residues

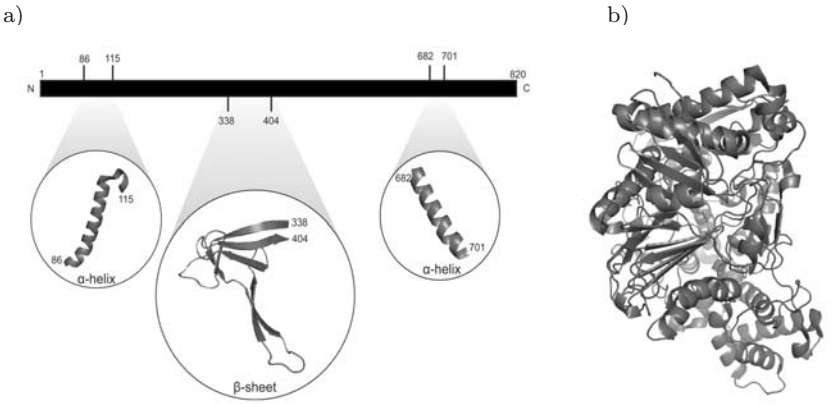

Fig. 3.15 Protein structure exemplified by pro-penicillin amidase (protein data bank, PDB, code 1E3A): **a** The primary amino acid amino acid sequence can form different secondary structures; **b** The backbone of the 3D structure. The crystal structure of pro-penicillin amidase was solved by L. Hewitt and coworkers (2000).

(π-turn). The Ω-loop is a term for a longer loop (named because of its similarity to the Greek upper case letter omega) comprising 6 to 16 residues with an end-to-end distance of less than 10 Å. The Ω-loop also adopts a compact conformation, in which the side-chains tend to fill the internal loop cavity.

The secondary structures are spatially ordered to one another in the 3D space forming the *tertiary structure*. Non-polar residues (e.g., Leu, Ile, Val, Phe) are sequestered in the interior of the molecule in the globular proteins, thus avoiding the contact with water from the environment. In turn, the proteins' surfaces are enriched with charged polar residues (e.g., Arg, Lys, His, Asp, Glu) that can establish contacts with the aqueous exterior. In special cases, these residues can be found in the interior of the protein, where they accomplish specific functions (e.g., catalytic functions in the enzymes). Uncharged polar residues (e.g., Ser, Asn, Gln, Tyr) can be found both on the surface and in the interior of the molecule. When found buried in the interior, they are almost always involved in hydrogen bonds. Tertiary structure (i.e., the long- or short-range interactions between the side chain of the amino acids), and the propensity to establish different secondary structures, are assumed to be determined by the primary sequence of the protein. Various secondary interactions between the backbone and different side chains provide the structural basis for the native 3D pattern of a protein sequence: (1) electrostatic interactions, including dipole-dipole and ionic interactions, (2) hydrogen bonds, (3) hydrophobic interactions, and (4) covalent disulfide bonds, formed between some free sulfhydril groups in the side chains of the cysteines in an oxidizing environment. The tertiary structural arrangement of the proteins ensures dense packing with a minimized ratio of the volume enclosed in van der Waals interactions (Fig. 3.15). This ratio of about 0.75 is within the same range of crystals formed from small organic molecules.

The 3D structure, in which a protein can accomplish its physiological functions in the cell, is also called native fold or native conformation. It is commonly assumed that in the cellular environment, the native conformation is the most thermodynamically stable conformation, populating the global minimum of the energy. En route to the stable native fold, some proteins go through several partially folded intermediate states, which dwell in local energy minima. In the cell, a variety of other proteins (e.g., chaperones, disulfide isomerases, catalyzing the formation of disulfide bonds) assist the newly synthesized proteins in attaining their native conformation. For some proteins, the dwelling in a local energy minimum might have a physiological role. A classical example is the hemagglutinin, in which the two chains of the mature protein are kinetically trapped in an intermediate state. A drop in the pH causes conformational changes in the intermediate to an energetically more favorable state, which enables it to penetrate the host cell membrane. Proteins are not rigid molecules and might shuttle between different struc-

tures or conformations, which allows them to accomplish their physiological functions (e.g., binding of substrates to the active site of enzymes).

Several polypeptide chains that have independently acquired a 3D structure can interact in order to attain a highly ordered structure, a *quaternary structure*. The individual chains are called subunits and they are stabilized by the same type of interactions, responsible for the stability of the 3D structure (i.e., secondary non-covalent interactions or disulfide bonds within the protein complex). In one such complex two (dimer), three (trimer) or more polypeptides (multimer) can be associated. Proteins consisting of identical subunits only are referred to with a prefix "homo-" (homodimer, homotrimer, etc.), whereas the complexes containing structurally different subunits are assigned the prefix "hetero-" (heterodimer, heterotrimer, etc.).

3.4.3 Enzymes

Enzymes are proteins that accomplish specific catalytic functions in the cell. They enormously accelerate chemical reactions in the cell (10^6–10^{12} increase of the rate over the corresponding uncatalyzed reaction). Enzymes are specific and act as catalysts only in one or a few similar reactions and are involved in all biochemical reactions in the cell (e.g., DNA replication, RNA transcription, catabolic (degradation) and anabolic (synthesis) reactions). A small fraction of the amino acids of the entire enzyme molecule form the catalytic, active center that comes into contact with the *substrate*, a molecule that will be converted to one or more *products*. This segment, usually consisting of three or four residues, is called the *active site*. Similar to the catalytic mechanism of the chemical catalysts, the enzymes lower the activation energy of the reaction (E_a or $\Delta G^{\ddagger}$), thus accelerating the rate of reaction; they do not change the reaction pathway and do not alter the equilibrium (Fig. 3.16). Unlike the chemical catalysts the enzymes rarely produce side products and show a significant level of stereospecificity (recognize only one stereoisomer as a substrate) and regioselectivity (specific for only one substrate or small range of related compounds). The most specific enzymes are involved in the amplification and storage of the genome information (e.g., DNA polymerase, RNA polymerase, ribosomes, aminoacyl-tRNA synthase). Mammalian polymerases have an extreme fidelity with an error rate of about 10^{-7}. Some enzymes in the secondary metabolic pathways are described as promiscuous, because they can act on a broad range of different (but structurally similar) substrates.

The substrate binds to the enzyme through geometrically and physically complementary interactions and the binding is controlled through non-covalent forces which are identical with the interactions stabilizing different protein conformations. The first theory of the enzyme-substrate

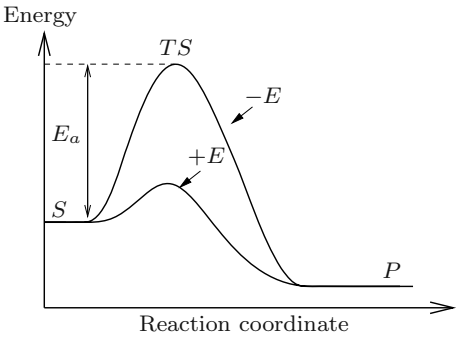

Fig. 3.16 Energy landscape of enzyme-catalyzed ($+E$) and uncatalyzed ($-E$) conversion of substrate (S) into product (P). When processing from higher energy S to lower energy P, the substrate must first pass the transition state TS, and the energy to reach the activation energy E_a represents the energy barrier of the reaction.

binding was developed by E. Fischer in 1894 suggesting that both molecules fit to each other based on a geometrically similar shape (e.g., "the lock and key" model). Although specificity is well covered by this model, it cannot explain the catalytic mechanism. Koshland's theory (1958), known as *induced fit*, suggests that substrate and enzyme are both flexible molecules and interactions between them can reshape their conformational state.

Some enzymes require other factors, *cofactors*, to be bound to accomplish their activity. Cofactors can be either inorganic (metal ions) or small organic molecules (heme, flavin, NADH, vitamins). Cofactors that remain tightly bound to the enzyme, whose function they assist, are known as *prosthetic groups* (NADH). Enzymes with a bound cofactor are called holoenzymes (i.e., active form), and enzymes without a cofactor apoenzymes. Co-enzymes are chemically changed during the enzyme reaction and regenerated at the end, thus maintaining constant steady-state levels inside the cell.

Enzyme activity can also be affected by other molecules, called *effectors*, which modulate the enzyme activity either by direct binding to binding sites of the enzyme or indirectly through subunits that modulate the enzyme function. Such enzymes are called *allosteric*. Small molecules can also decrease or inhibit enzyme activity (i.e., inhibitors), or enhance the enzyme activity (i.e., activators). Enzymes, like all the proteins, are conformationally dynamic and this internal dynamics can be essential for the catalytic properties. Temperature and environmental pH can influence the conformational freedom of the enzyme molecule and change their catalytic properties. In the cell, the enzyme activity is controlled additionally by its amount, which depends on the rate of biosynthesis and the rate of degradation.

Enzyme Kinetics

The conversion of a substrate S to a product P catalyzed by an enzyme E can be described by the chemical equation

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} E + P \; . \tag{3.38}$$

This process undergoes two stages. First, the substrate reversibly binds to the enzyme and an enzyme-substrate complex ES is formed. Second, the enzyme catalyzes the chemical step in the reaction and releases a product. The enzyme is not altered by the reaction and thus the equilibrium is influenced only by the thermodynamical properties of S and P. In equilibrium, the conversion to the product becomes rate-limiting. Therefore, the corresponding differential reaction-rate equation has the property

$$\frac{d[ES]}{dt} = k_1[E][S] - k_{-1}[ES] - k_2[ES] = 0 \; . \tag{3.39}$$

Hence,

$$[ES] = \frac{k_1[E][S]}{k_{-1} + k_2} \; . \tag{3.40}$$

In 1913, L. Michaelis and M. Menten assumed that product formation is the rate-limiting step ($k_{-1} \gg k_2$). The maintenance of the ES complex in equilibrium (steady-state assumption) is then described by the *Michaelis constant*

$$K_m = \frac{k_{-1} + k_2}{k_1} \approx \frac{k_{-1}}{k_1} \; . \tag{3.41}$$

The non-covalent ES complex is known as the Michaelis complex. Eq. (3.40) then simplifies to

$$[ES] = \frac{[E][S]}{K_m} \; . \tag{3.42}$$

The total concentration of the enzyme is the sum of the concentrations of the enzyme in the ES complex and the free soluble enzyme,

$$[E_0] = [E] + [ES] \; . \tag{3.43}$$

Thus, Eq. (3.42) rearranges to

$$[ES] = \frac{([E_0] - [ES])[S]}{K_m} \; , \tag{3.44}$$

which can further be transformed to

$$[ES] = [E_0] \frac{1}{1 + K_m/[S]} \; . \tag{3.45}$$

The initial velocity of the reaction is given by

$$\frac{d[P]}{dt} = k_2[ES] \; . \tag{3.46}$$

Substituting Eq. (3.45) into Eq. (3.46) and multiplying both nominator and denominator by $[S]$ gives

$$\frac{d[P]}{dt} = k_2[E_0]\frac{[S]}{K_m + [S]} = V_{max}\frac{[S]}{K_m + [S]} \tag{3.47}$$

where V_{max} is the maximum velocity. In this case, all enzyme molecules are saturated by the substrate and the enzyme exists only in an ES form:

$$V_{max} = k_2[E_0] \; . \tag{3.48}$$

k_2 is called *catalytic constant* or *turnover number* k_{cat} and gives the number of conversions (turnovers) that each catalytic site catalyzes per unit of time. Combining Eqs. (3.47) and (3.48), an equation known as the *Michaelis-Menten equation* is derived,

$$\nu = \frac{V_{max}[S]}{K_m + [S]} \; . \tag{3.49}$$

At the substrate concentration $[S] = K_m$, the reaction velocity is half-maximal. The Michaelis constant K_m is the measure of the affinity of the enzyme to the substrate: For enzymes with small K_m values, the maximal catalytic activity is achieved at low substrate concentrations. Each enzyme has a characteristic K_m value for a given substrate; the K_m value varies for different substrates of the same enzyme and it is a function of the temperature and pH value. The apparent second-order rate constant k_{cat}/K_m, known also as a *specificity constant*, is used as a measure of the catalytic efficiency of the enzyme. It depends on the encounters between substrate and enzyme in a solution and summarizes both characteristics of an enzyme: affinity and catalytic ability. The specificity constant can be used for the comparison of enzymes with different substrates. The theoretical maximum of k_{cat}/K_m is 10^8–10^9 per Ms and is called diffusion limit (i.e., the diffusion rate limits the reaction rate and every collision of the enzyme with a substrate will release a product).

Michaelis-Menten kinetics assumes irreversibility of the enzymatic catalysis and it is based on the law of mass action, assuming free diffusion and random collision. In the cell, however, the processes can deviate from such idealized conditions. The highly crowded internal space of the cell significantly limits the free molecular movements (e.g., the concentration of macromolecules in the cytoplasm of prokaryotic cells is about 400 mg/ml). For some heterogeneous processes (e.g., as in the case of DNA polymerase), the substrate mobility may also be limited. These deviations from the conventional

mass-action laws led to the development of the limited-mobility derived or fractal-like kinetics, reviewed by R. Kopelman (1988), M. Savageau (1995), and S. Schnell and T. Turner (2004).

Enzyme Thermodynamics

The enzymatic reaction (3.38) can be described by the equilibrium equation

$$-RT \log K_{eq} = \Delta G^{\circ}_{eq}, \tag{3.50}$$

where K_{eq} is the equilibrium constant, ΔG°_{eq} is the Gibbs free energy of the ES complex, T is the absolute temperature, and R is the gas constant. But if it is assumed that the formation of the ES complex is in rapid equilibrium with the reactants, then the equilibrium constant can be expressed as

$$K_{eq} = \frac{[ES]}{[E]\,[S]} \,. \tag{3.51}$$

Combining Eqs. (3.46), (3.50), and (3.51) gives

$$\frac{d[P]}{dt} = k_2[E][S]e^{-\Delta G^{\circ}_{eq}/RT} \,. \tag{3.52}$$

This equation shows that the rate of the reaction depends not only on the reactants, but increases exponentially with ΔG°_{eq}.

3.5 Cells and Organisms

Despite the phenotypic and genotypic variation of different organisms, they share remarkable similarities on the cellular level (e.g., in the general cell structure, physiology and biochemistry). The cells propagate by duplicating the DNA and each daughter cell inherits identical genetic material from the parental cell. This genetic information is translated in a variety of proteins that determine the functional diversity of each cell. The prokaryotic cells are commonly unicellular and a single cell is capable of executing all the physiological activities. Unlike the prokaryotes, in the eukaryotic cell the physiological activities are spatially separated in compartments or organelles surrounded by double-lipid-layer membranes. The various membrane-separated organelles in the cell that accomplish different functions are referred to as the *endomembrane system*. Another level of complexity exists in multicellular eukaryotic organisms: various physiological activities are separated in different cells and have led to cell specialization and differentiation of the cell types.

3.5.1 Eukaryotes and Prokaryotes

The nucleus is the largest organelle in the eukaryotic cell and carries the inheritance information encoded in the DNA. It is surrounded by a double membrane, referred to as the *nuclear envelope*, with pores that allow nucleus components to move in and out. The nuclear membrane extends in various tube- and sheet-like membrane extensions, forming an extensive network of membranes, called the *endoplasmic reticulum*. In eukaryotes, the mRNA is synthesized (transcription) in the nucleus, whereas the translation into proteins occurs in the cytosol. Ribosomes that perform the protein synthesis are attached to the endoplasmic reticulum. Furthermore, the newly synthesized proteins are modified and transported to their final destinations in the *Golgi-bodies* which bud off from the endoplasmic reticulum. The endoplasmic reticulum and Golgi apparatus function not only in the processing and transport of proteins designed for secretion and membrane incorporation, but also actively participate in the synthesis of lipids. Different vesicles can be formed by budding off the membranes whose function is to transport nutrients into and waste products out of the cell.

In nearly all eukaryotes the aerobic respiratory functions are accomplished in the mitochondria, which are surrounded by a double layer of membranes. Mitochondria are found in almost all eukaryotes and play a critical role in energy metabolism. They produce the energy substance, ATP, by oxidative breakdown of the nutrients. Mitochondria contain their own DNA. There is a variety of other simple compartments surrounded by membranes that are responsible for different functions found in the eukaryotic cell:

- Lysosomes contain proteolytic enzymes that digest the food substances.
- Peroxisomes allow specific reactions in a defined environment to take place in which toxic peroxide is released.
- Vacuoles maintain the osmotic pressure of the higher plants.
- Chloroplasts produce energy through photosynthesis and are characteristic of plant cells and various groups of algae.

Next to the spatial separation of biochemical processes, the eukaryotic cell has another level of organization: a network of filamentous proteins throughout the cytoplasm forms the *cytoskeleton*, which provides the cell cytoplasm with structure and shape. The cytoskeleton determines the general organization within the cell and enables motions of the entire cell and throughout the cytoplasm. The cytoskeleton is composed mainly of actin filaments, intermediate filaments, and microtubules.

Prokaryotes are referred to as organisms whose genetic information is not stored in the nucleus or any membrane-bound structure. They are usually smaller and their cell structure is simpler than that of eukaryotes. The genetic material is encoded by a single molecule of chromosomal DNA called also *nucleoid*. Some prokaryotic cells might contain self-replicating satellite circular DNA, known as *plasmids*, which encodes some crucial functions for the

prokaryotic cell. The prokaryotes also lack several membrane-bound cell compartments; nevertheless, the biochemical processes are spatially separated within the cell, positioned at differents subcellular sites. Cytoskeleton elements, homologous to eukaryotic cytoskeletal proteins, have also been found in prokaryotes and they function in actively positioning proteins and DNA molecules. Prokaryotes have a shorter generation time and a large surface-to-volume ratio that consequently gives them a higher metabolic rate compared to eukaryotes. Based on the ability of prokaryotes to utilize oxygen or different compounds for oxidation processes, the cells are divided into aerob (use oxygen) and anaerob.

3.6 Viruses

Viruses are sub-microscopic particles that can infect the cells of living organisms. They are not considered as eukaryotes or prokaryotes.

3.6.1 General Structure and Classification

Viruses (translated from Latin as toxin or poison), also known as *virions*, infect both eukaryotic and prokaryotic cells and replicate and propagate further using the transcription and translation systems of the host cell. The group of viruses infecting bacteria is known as bacteriophages. The majority of the viruses have a size of 10–250 nm and with some exceptions up to 750 nm, which is larger than the size of some bacteria. The genetic information is stored on nucleic acid, which can be either RNA (single- or double-stranded) or DNA (single- or double-stranded), and is encapsulated by a protein shell, known as *capsid*. The proteins of the capsid are encoded by the viral genome. Viral particles have regular shapes including helical capsids (tobacco mosaic virus), icosahedral symmetry (hollow quasi-spherical structure; polio virus, and foot and mouth disease virus), enveloped viruses (in addition to the capsid, they are covered by a lipid-bilayer membrane; influenza virus and human immunodeficiency virus), and complex viruses (tailed bacteriophages, poxviruses). Some viruses are unable to survive outside the host cell, whereas others are more stable and can persist outside a cellular environment for very long periods.

According to the classification proposed by the Nobel Prize-winner D. Baltimore, also known as Baltimore classification, viruses can be divided into seven groups: (I) double-stranded DNA, (II) single-stranded DNA, (III) double-stranded RNA, (IV) positive-sense (+) single-stranded RNA, (V) negative-sense (−) single-stranded RNA, (VI) single-stranded-RNA reverse transcribing, and (VII) double-stranded-DNA-reverse transcribing. In the DNA viruses (Groups I and II), the genetic information is stored on DNA

and it is replicated via DNA-dependent DNA polymerase, while in the viruses belonging to Groups III, IV, and V the genetic information is stored in an RNA sequence. The ribosomes of the host directly translate the RNA strand into the positive-sense viruses, whereas in the negative-sense viruses the RNA is first inverted into positive-sense RNA via RNA polymerase. Examples of RNA viruses are hepatitis A and C, SARS, yellow fever, rubella and influenza viruses; to the DNA viruses belong simian virus (SV) 40, human herpes virus, cowpox smallpox viruses, and bacteriophages. The most severe Marburg, Ebola, and Lassa viruses belong to the group of negative-sense (−) single-stranded RNA viruses.

In the reverse transcribing single-stranded RNA viruses (Group VI), also called retroviruses, the genetic information is stored on RNA, and they replicate by formation of DNA due to reverse transcription via RNA-dependent DNA polymerase. The DNA is then integrated into the host genome and further propagated by the replication and translation machinery of the host cell. The most prominent member of this group is the human immunodeficiency virus (HIV). The double-stranded DNA of the reverse transcribing double-stranded DNA viruses (Group VII) is transcribed both into mRNA and translated further into protein and RNA, which is integrated into the host genome after reverse transcription into DNA. To the latter Group belongs the hepatitis B virus.

Therapeutical approaches against viral diseases include vaccination or administration of anti-genic material to trigger immunity responses to the disease agent. Antibiotics and other drugs are often applied too, although their targets are mainly inflammation and other secondary responses caused by the viral infection.

3.6.2 Applications

Viruses have a great potential in the virotherapy also called oncolytic or viral therapy. This technique involves specific targeting and killing of cancer cells through the introduction into the body of genetically modified viral material. The viral particles reproduce rapidly over a short period of time and attack only the cancerous cells without harming the healthy cells. In the viruses tailored for the viral therapy, either the protein coat is modified so that the affinity only against cancer cells is assured, or the genetic information of the virus is altered and it can enter any cell but replicate only in the cancerous cells. This therapy can be applied to treat cancer as well as to inhibit angiogenesis. A critical barrier in the widespread application of cancer treatment is the immune system of the host, which responds to the engineered viruses and destroys them.

The shape and size of the viruses, the functional groups on their surface, and the tools they have developed to cross barriers of the host cells make

viruses very attractive to material sciences and nanotechnology. The precisely defined pattern of functional groups and their ordering on the virus surface offers a unique scaffold for covalently linked surface modifications. A. Belcher and colleagues (2002) have engineered a liquid crystal system from genetically engineered bacteriophage and zinc sulfide, which spontaneously assembles into a thin hybrid nanocrystal ordered into approximately 72 μm domains. In a subsequent study (2006), they have extended the application of the virus-templated synthesis to assemble nanowires of hybrid gold-cobalt oxide. The virus particles were genetically modified to incorporate gold-binding peptides into the filament coat. The total negative charge of these particles allows them to be layered between oppositely charged polymers to form thin and flexible sheets. This two-dimensional assembly of viruses on polyelectrolyte multilayers is a step forward in creating flexible ion batteries that supply much electrical energy in a thin and lightweight package.

HIV

Human immunodeficiency virus type 1 (HIV-1) is the etiologic agent of the immune deficiency syndrome (AIDS) and its related disorders. The first case report of AIDS appeared in 1981, while the virus was first isolated in 1983. Currently, there are about 49,000 HIV-1 infections in Germany alone, with about 2,900 new infections per year. Primarily, HIV-1 infects and kills immune cells which regulate and amplify immune response. Without effective anti-retroviral therapy, the hallmark decrease in immune cells during AIDS results in a weakened immune system that impairs the ability to fight against infections or cancer, so that death eventually results. HIV-1 is a retrovirus, and similar to all retroviruses its RNA genome replicates by a DNA intermediate. Many retroviruses contain three structural genes (Gag, Pol, and Env, encoding for core and structural proteins, reverse transcriptase, and coat proteins, respectively). HIV-1 has additional accessory and regulatory genes: Vif, Vpr, Vpu (accessory) and Nef, Tat, Rev (regulatory).

The HIV-1 life cycle is well-documented. HIV infection begins when the viral glycoprotein (gp) interacts with the CD4 receptor on the surface of an immune cell. After fusion of the viral membrane with the cell membrane, the viral core with the associated RNA is internalized into the cell. Partial uncoating of the viral core exposes the viral RNA. In the cytoplasm of the recipient cell, the viral genomic RNA is synthesized by viral reverse transcriptase into a viral double-stranded DNA preintegration complex. After migrating into the nucleus, facilitated by other viral proteins the viral double-stranded DNA of the preintegration complex is integrated randomly into the host DNA by viral integrase. RNA polymerase transcribes the proviral DNA into mRNA. In the early transcription phase, the mRNA is spliced by the cellular splicing machinery into multiply spliced transcripts, producing the Tat, Rev and Nef proteins. When Rev accumulates to a critical level, the mRNA production shifts from multiple spliced to single spliced and unspliced transcripts,

resulting in the Env gp160 (containing envelope proteins gp120 and gp41), Vif, Vpr, Vpu proteins and Gag p55 (containing matrix, capsid, nucleocapsid) and Gag-Pol p160 (containing matrix, capsid, protease, reverse transcriptase, and integrase) proteins, respectively. The virus is assembled at the plasma membrane, forming a budding virion. After virus budding out from the cell surface into extracellular space, maturation of the virus proceeds (by proteolytical cleavage of the Gag and Gag-Pol polyprotein). Now the virus is ready for another round of infection.

Hepatitis B Virus

Hepatitis B virus (HBV) has one of the smallest genomes (approximately 3 kb) and belongs to the group of reverse transcribing double-stranded DNA viruses. The virus specifically infects the liver of humans and various animals and causes acute liver damage. Over 250 million people worldwide are infected with HBV yearly, some of whom develop severe pathological consequences, including chronic hepatitis, cirrhosis, and hepatocellular cancer. HBV infection is particularly common in Asia and Africa and is associated with approximately 10% of the worldwide liver cancer incidence (up to a million cases of liver cancer annually). Transfection occurs through infected blood and other body fluids.

The hepatitis B virus is composed of an outer lipid envelope and an icosahedral nucleocapsid core (i.e., DNA genome and protein coat surrounding it). The virus attacks the surface receptors of the hepatocytes and after internalization into the cell migrates into the nucleus. It replicates through reverse transcription of RNA intermediate, the pregenomic RNA, which can be packed into capsids and is reversely transcribed into DNA. The new virions bud out of the endoplasmic reticulum and are exported from the cell.

In the cases of acute HBV infection, up to 95% of adults overcome the infection spontaneously without any treatment. Chronic HBV is treated with anti-viral agents that either inhibit the virus replication or stimulate the immune response. Five drugs have been approved for the treatment of chronic HBV infection: interferon-alpha, pegylated interferon-alpha, lamivudine, adefovir dipivoxil, and entecavir. Their efficacy is limited by their side effects, as well the high frequency of viral mutations which render the therapeutics less potent.

References

1. Ambrogelly A, Palioura S, Soll D (2007) Natural expansion of the genetic code. Nat Chem Biol 3:29–35
2. Bock A, Forchhammer K, Heider J, Baron C (1991) Selenoportein synthesis: an expansion of the genetic code. Trends Biochem Sci 16:463–467
3. Breslauer KJ, Frank R, Blöcker H, Marky LA (1986) Proc Natl Acad Sci 83: 3746–3750

4. Brett D, Pospisil H, Valcárcel J, Reich J, Bork P (2001) Alternative splicing and genome complexity. Nature Gen 30: 29–30
5. Bundschuh R, Gerland U (2006) Dynamics of intramolecular recognition: base-pairing is DNA/RNA near and far from equilibrium. Eur Phys J 10:319–329
6. Coffin JM (1996). Retroviridae. In: Fields BN, Knipte DM, Howley PM (eds.) Fields Virology Raven Publ
7. Collier J, Shapiro L (2007) Spatial complexity and control of a bacterial cell cycle. Curr Opin Biotechnol 18:333-340
8. Gibson M, Bruck J (2000) Efficient exact stochastic simulation of chemical systems with many species and many channels. J Phys Chem A 104:1876–1889
9. Gillespie D (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J Phys Chem 22:403–434
10. Gillespie D (1977) Exact stochastic simulation of coupled chemical reactions. J Phys Chem 81:2340–2361
11. Hewitt L, Kasche V, Lummer K, Lewist RJ, Murshudov GN, Verma GN (2000) Structure of a slow processing precursor penicillin acylase from Escherichia coli reveals the linker peptide blocking the active cleft. J Mol Biol 302:887–898
12. Hohsaka T, Sisido M (2002). Incorporation of non-natural amino acids into proteins. Curr Opin Chem Biol 6:809–815
13. Jimenez-Sanchez A (1995) On the origin and evolution of the genetic code. J Mol Evol 41:712–716
14. Kelly E, Russell SJ (2007) History of oncolytic viruses: genesis to genetic engineering. Mol Ther 15:651–659
15. Kopelman R (1988) Fractal reaction kinetics. Science 241:1620–1626
16. Lee SW, Mao C, Flynn CE, Belcher AM (2002) Ordering of quantum dots using genetically engineered viruses. Science 296:892–895
17. Mattick JS, Makunin IV (2006) Non-coding RNA, Hum Mol Gen 15:17–29
18. McMahon BJ (2005) Epidemiology and natural history of hepatitis B. Semin Liver Dis 25:3–8
19. Nam KT, Kim DW, Yoo PJ, Chiang CY, Meethong N, Hammond PT, Chiang YM, Belcher AM (2006) Virus-enabled synthesis and assembly of nanowires for lithium ion battery electrodes. Science 312:885–888
20. Noad R, Roy P (2003) Virus-like particles as immunogens. Trends Microbiol 11:438–444
21. Ryu WS (2003) Molecular aspects of hepatitis B viral infection and the viral carcinogenesis. J Biochem Mol Biol 36:138–143
22. SantaLuica J Jr, Hicks D (2004) The thermodynamics of DNA structural motifs. Annu Rev Biomol Struct 33:415–440
23. Savageau MA (1995) Michaelis-Menten mechanism reconsidered: implications of fractal kinetics. J Theor Biol 176:115–24
24. Schnell S, Turner TE (2004) Reaction kinetics in intracellular environments with macromolecular crowding: simulations and rate laws. Prog Biophys Mol Biol 85:235–260
25. Tijssen K (1993) Laboratory techniques in biochemistry and molecular biology. Elsevier, Amsterdam
26. Varani G, McClain W (2000) The `G×U` wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. EMBO Rep 1:18–23
27. Vaha-Koskela MJ, Heikkila JE, Hinkkanen AE (2007) Oncolytic viruses in cancer therapy. Cancer Lett 254:178–216
28. Wagner E (2007) Programmed drug delivery: nanosystems for tumor targeting. Expert Opin Biol Ther 7:587–593
29. Wilkinson DJ (2006) Stochastic modelling for systems biology. Chapman & Hall, New York
30. Xie J, Schultz PG (2006) A chemical toolkit for proteins – an expanded genetic code. Nat Rev Mol Cell Bio 7:775–782