Appendix

Principles of Protein Structure and Function

A.1 Brief Introduction to Structural Cell Biology

During the last century, the life sciences have witnessed an enormous advance in methodology and technology which enabled scientists to perform demanding experiments and to get scientific data and results of previously unknown quality and quantity. One of the most important steps in the life sciences during the last decade was the successful sequencing of the human genome which enabled scientists from many scientific disciplines to analyze genomic data from our own species. These achievements however, raised more questions than they could provide answers concerning the properties of complex multicellular organisms and the molecular nature of life in general.

One of the most surprising results of the human genome project was the estimated total number of genes in the human genome (ca. 3,000,000 base pairs), which turned out to be in the range of only 20,000 – 25,000 protein-coding genes. Surprisingly, there is no clear correlation between the number of genes and the nuclear genome size (known as ***C-value paradox***). In fact the human genome is expected to consist largely of non-coding or regulatory DNA and only 1.5% protein-coding sequences. This may be considered as strong indication that it needs much more than the sequence of the genome to understand what makes the difference between man, monkey and yeast. Although complex mechanisms of regulation on the genomic level are known, it is now widely accepted that an organism's complexity arises from the structural and functional links between different proteins and other gene products which allow an almost unlimited number of precisely regulated molecular interactions. A deeper understanding of the molecular nature of life therefore requires knowledge about the physical and chemical properties of the molecules present in living cells. In addition, it is not only necessary to know which players are potentially in the field, but it is equally important to get a picture of the different kinds of molecular interactions within living cells and the regulation of these events with respect to intra- and extracellular signals or stimuli.

By far the most abundant and important gene products are proteins which display an enormous diversity in terms of sequence, structure and function. Protein activity itself is regulated by complex biochemical mechanisms, including all levels from protein expression to degradation. To fully understand these complex molecular events, scientists aim at a comprehensive description of the structure and function for all proteins of a given species. This research field has been termed ***proteomics***. Proteins must interact with matter (i.e. other molecules or solutes in the cell) or at least with energy (i.e. light, or more general: electromagnetic radiation) if they are to exert a certain cellular function whatsoever, and it is the defined three-dimensional architecture that enables proteins to fulfill their biological functions. The combination of structural and functional data is therefore extremely valuable for the characterization of proteins and a deeper understanding of biochemical mechanisms in living cells. In recent years, previously more or less isolated disciplines in life sciences have begun to approach and to benefit from the interdisciplinary research with scientists from the fields of structural biology, biophysics, biochemistry and medicine involved.

A.2 From Sequence to Structure: Protein Folding

Gene expression in terms of transferring genetic information from genes to functional proteins involves three fundamental steps: (i) transcription of DNA into mRNA, (ii) translation of mRNA into protein, and (iii) protein folding. For many proteins, post-translational modifications may occur to regulate function and activity. To exert a biochemical function whatsoever, the linear one-dimensional information associated with the amino acid sequence of polypeptides needs to be converted into functional proteins. This step from an unstructured polypeptide chain to a three-dimensional protein with defined structure is known as protein folding. In chemical terms, protein folding may be considered a chemical reaction, $P_{\mathbf{unfolded}} \rightarrow P_{\mathbf{folded}}$, with a change in free energy (ΔG) given as:

$$\Delta G = \Delta H - T\Delta S = -RT \bullet \ln(K)$$

with

$$K = [P_{\mathbf{folded}}] / [P_{\mathbf{unfolded}}]$$

ΔH and ΔS denote the change of enthalpy and entropy, respectively, $[P_{\mathbf{folded}}]$ and $[P_{\mathbf{unfolded}}]$ are the chemical activities (essentially the concentrations) of folded and unfolded polypeptide, respectively. R is the (ideal) gas constant and T is the absolute temperature in Kelvin. As

shown by the formalism above, protein folding suggests a trade-off between enthalpy and entropy. The conformational entropy of the polypeptide in the folded state is usually lower than that in the unfolded state, whereas the binding energy normally increases in the course of folding as the native conformation is stabilized by hydrogen bonds, ionic and hydrophobic interactions. A comprehensive description of protein folding, however, requires the entropy change of both the polypeptide *and* the surrounding aqueous solution to be accounted for. In simple terms, protein folding aims at sequestering hydrophobic segments from the aqueous surroundings by burying them in the interior of the folded protein. The driving force behind is known as ***hydrophobic effect*** which says that it is favorable for energetic reasons to assemble hydrophobic amino acid residues in the inner core of the folded polypeptide, and to expose hydrophilic amino acid residues to the surrounding water molecules. By this, the solvent accessible surface becomes effectively reduced in size and additionally restricted to residues with a high tendency of forming hydrogen bonds. Yet the total number of hydrogen bonds between hydrophilic amino acid residues and solvent molecules is largely unchanged or may become even reduced in the folding process as *inter*molecular hydrogen bonds to solvent water molecules may become replaced by *intra*molecular interactions in the folded state. Solvent exposed hydrophobic segments, however, cause water molecules to form cage-like clathrate structures which results in a decrease of the solvent entropy with only minor energetic (i.e. enthalpic) compensation. Thus, burying hydrophobic amino acid residues in the protein inner core may provide a substantial pay-back in entropy. Considering the folding polypeptide plus solvent molecules as thermodynamic system, the *overall* change of both entropy (ΔS) and enthalpy (ΔH) for the folding reaction $P_{\mathbf{unfolded}} \rightarrow P_{\mathbf{folded}}$ may be, however, rather small, as may be the net change in free energy ($\Delta G = \Delta H - T\Delta S$), the latter being typically in the range of 5-15 kJ/mol. Thus, the native conformation of a folded protein is only marginally stable with respect to the unfolded state.

Although the underlying principles are well understood, the process of protein folding has puzzled scientists over decades and still represents one of the most complex reactions in biochemistry. This is due to the extremely large numbers of possible intermediate states, creating a highly complex n-dimensional conformational space with n being the number of degrees of freedom. If the search for the native conformation were to proceed randomly, the process of protein folding would require more time than the age of the universe. Protein folding *in vivo*, however, is mostly completed within a few seconds. This apparent discrepancy is known as the ***Levinthal paradox*** [71]. Therefore, protein folding cannot be completely random but is apparently governed by a driving force that effectively narrows the conformational space. To be thermodynamically stable, folded proteins need to populate lower free

energy levels than unfolded polypeptides. If this were not the case, the equilibrium between folded and unfolded states would be shifted far to the latter and protein folding would be extremely inefficient.

A.2.1 Energy Landscapes and Folding Pathways

To solve the Levinthal paradox, the concept of ***energy landscapes*** and ***folding pathways*** has been developed [72,73]. Since every possible polypeptide conformation is associated with a certain free energy level, protein folding may be considered a series of conformational changes along specific trajectories (folding pathways) with a step-wise decrease of free energy down to the single native conformation. The different intermediate states in the course of protein folding can be visualized by means of energy landscapes which correlate different conformations with the associated free energies. For pictorial representation, the energy landscape may be considered three-dimensional, depicting two conformational degrees of freedom over the free energy. This might be used, for instance, to correlate the two backbone angles (phi and psi) with the free energy of conformation for a dipeptide. The resulting energy landscape has a funnel-like shape, with the point of lowest energy at the bottom representing the native conformation (N). The area size enclosed by the energy landscape at a given free energy level (i.e. the width of the funnel) is proportional to the number of possible conformations and therefore reflects the conformational entropy of the system at the chosen energy level (Figure series 16).

The Levinthal paradox refers to the many possible conformations a folding polypeptide chain needs to explore on the search for its native conformation. This situation is represented by an energy landscape with a geometry resembling that of a golf-course: a totally flat surface with a single deep energy minimum corresponding to the native state (Figure 16a). In this case, the free energy levels of non-native states are all-equal, thus folding follows no specific path but is instead entirely random. The surface area of the energy landscape reflects the conformational entropy of the system which is, in case of the golf-course landscape, enormously large, thus the folding reaction becomes extremely slow. However, the conformational space to be explored in the course of the folding reaction may be substantially reduced by a pathway that essentially restricts the number of accessible conformations, thus guiding the polypeptide from a denatured state (A) to the native state (N) (Figure 16b). As a result, the search is much more directed (i.e. energetically biased) toward the native conformation, and folding is substantially faster.

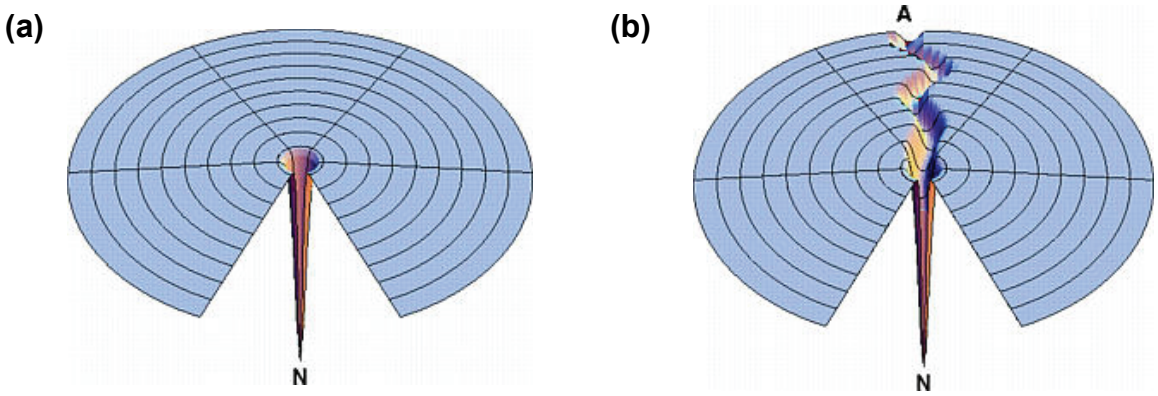

Figure series 16 (I): Energy Landscapes and Protein Folding Funnels.
(a) The Levinthal "golf-course" landscape. The flat surface is depictive of the all-equal free energy level of non-native conformations. Thus, the search for the native conformation (N) is completely random. The surface area of the energy landscape represents the conformational space to be explored in the search for N, thus reflecting the system's entropy. **(b)** Pathway solution to the random search problem depicted in (a). Figures adopted from [72,73].

However, the energy levels of non-native peptide conformations are by no means all-equal but differ substantially. Thus the population of energy levels is not completely random as it would be in the case of an entirely flat energy landscape with all-even golf-course geometry. Instead protein folding is highly cooperative which implies that sequential steps of the protein folding reaction are coupled such that preceding intermediates strongly favor conformational changes leading to further improvement and productive folding. This effectively narrows the conformational space to be explored and channels the different folding trajectories down to the lowest-energy structure. The corresponding energy landscape is characterized by a funnel-shaped surface geometry (Figure 16c). An important point is that there is no single exclusive folding pathway for any protein with defined sequence. Depending on the starting point and the biophysical restraints, a polypeptide may successfully follow different trajectories on the free energy surface (Figure 16d). The simplest folding pathway is highly cooperative and runs along an ideally smooth surface (trajectory A). A more realistic depiction might include energy barriers and kinetic traps which eventually slow down the folding rate (trajectory B). Although the rate constants may differ, both of the different trajectories are representative of productive folding pathways. The entirely smooth folding funnel is, however, an unrealistic oversimplification. The funnel-shaped energy surface is in fact heavily rugged with local minima and maxima which may function as kinetic traps and energy barriers, respectively (Figure 16e). As a result, only a limited number of folding pathways are used by a defined polypeptide. By this, the number of conformations to be scanned for in the folding reaction to find the native state is reduced by orders of magnitudes and protein folding can be accomplished within seconds. This is basically the solution to Levinthal's paradox.

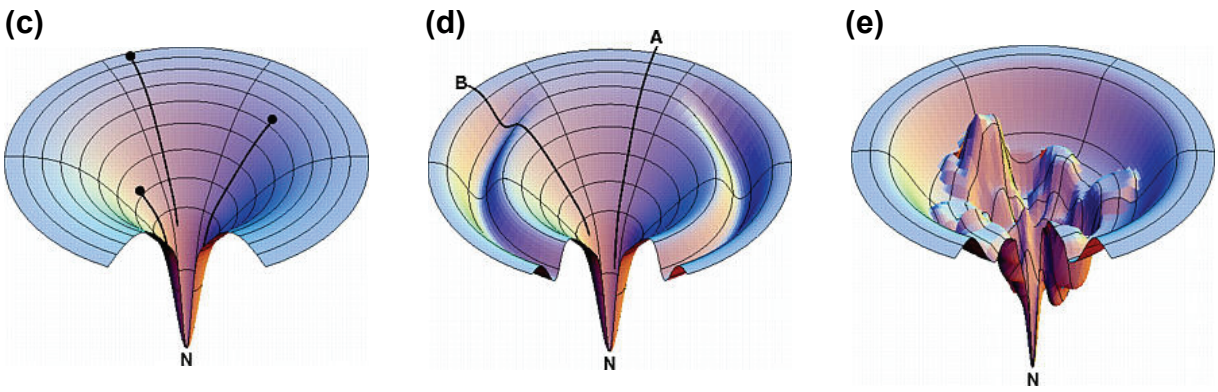

Figure series 16 (II): Energy Landscapes and Protein Folding Funnels.
(c) Idealized funnel-shaped energy landscape. The free energy surface is completely smooth and bears no kinetic or energetic restrictions. Thus, polypeptides with non-native conformations are effectively channeled along different folding trajectories or pathways (black lines) down to the single native structure (N). **(d)** Moat energy landscape, depicting two different folding pathways. Trajectory A represents a fast-folding process along an ideally super-smooth energy surface. Trajectory B includes a kinetic trap which is separated by an energy barrier from the native state. Note that both pathways lead to successful protein folding. **(e)** More realistic rugged folding funnel with kinetic traps and energetic barriers which force the protein folding reaction to proceed along narrow throughway pathways to the native state (N). Figures adopted from [72,73].

The concept of energy landscapes and folding funnels has important implications for the understanding of protein folding. Firstly, the polypeptide chain loses entropy during the folding process because the conformational space becomes more and more restricted. This is reflected in the narrow width of the funnel at low energy levels. Consequently, the conformation with the lowest free energy level has the lowest entropy, and the conformational space becomes restricted to a single main-chain conformation that is generally considered to represent the native structure (labeled N in Figure series 16). More precisely, the energy barrier between different side-chain rotamers or different conformations of unstructured regions may be very small, and proteins in solution are subjected to thermal motions and conformational fluctuations around the ideal single energy minimum. The ideal single-point energy level representing the native state is therefore modulated by a fine structure of free energy due to the dynamic behavior of the folded protein in solution. By using a quantitative approach, energy landscapes become (n+1)-dimensional, with n degrees of freedom. Although rather difficult to visualize in more than three dimensions, the concept of energy landscapes proved to be very helpful for a deeper understanding of the molecular mechanisms of protein folding and stability. Prion proteins appear to be the exception to the rule and are discussed below.

Exception to the Rule: Prion Proteins

In recent years prion proteins challenged the formalism of protein folding and the concept of energy landscapes which implied only a single native conformation to be possible.

Prion proteins (PrP) are localized to neurons and may adopt (at least) two different stable conformations with one of these, PrP^{sc}, being considered the molecular basis for scrapie and Creutzfeldt-Jakob disease. Unfortunately, PrP^{sc} appears to promote the PrP to PrP^{sc} transition. By this, PrP^{sc} is one of the rare examples for (naturally occurring) proteins that may adopt (at least) two different folds, both of which are stable under physiological conditions [74]. Other proteins that can adopt cross-β conformations, thus eventually forming amyloid-like deposits, are amyloid-β (Aβ) peptide, a proteolytic by-product of the transmembrane amyloid precursor protein (APP) which accumulates in the brains of patients with Alzheimer's disease (AD), α-synuclein, a (presumably) unstructured neuronal protein that is found in brains of patients with Parkinson's disease, and transthyretin (TTR) that is associated with neurodegenerative diseases such as familial amyloid cardiomyopathy (FAC) and familial amyloid polyneuropathy (FAP). The situation is reflected in an energy landscape with two (or more) distinct low-energy states, representing the biologically active conformation and distinct (kinetically) stable folding intermediates, separated by a substantial energy barrier. In this case, mutations may stabilize the non-native intermediate states, thereby altering the energy landscape in a way that the folding reaction comes entirely under kinetic control such that the transition between trap-intermediates and the biologically active conformation is effectively blocked and folding intermediates become kinetically trapped. Consequently, the distribution of conformations is no longer proportional to the associated free energy levels, although the relative free energy levels can be very similar. In the case of prion proteins the intermediate conformation related to PrP^{sc} shows a strong tendency toward aggregation into high-molecular complexes which may interfere with normal cellular functions.

Protein Folding *in vitro* and *in vivo*

There is a clear-cut difference between *in vitro* and *in vivo* experiments: ***in vitro*** means any situation in an *artificial environment* and is mostly used to describe the situation in the test tube, whereas ***in vivo*** refers to *living cells* or organisms. As such, *in vivo* experiments demand for the use of non-destructive methods to maintain the viability of the system. This may be accomplished by specific molecular labels such as green fluorescent protein (GFP) for various types of reporter assays. To investigate the cellular function of proteins or other molecular targets *in vivo*, genetically modified organisms (e.g. knock-out mice) may provide clues. *In vivo* experiments are demanding in terms of time and expertise, therefore by far most experiments are done in the test tube, i.e. *in vitro*. Some experiments aim at identifying intracellular interactions by means of standard test tube detection systems, such as antibody-based co-precipitation methods. In this case, the term ***ex vivo*** is occasionally used.

Although the fundamental principles stay the same, protein folding *in vitro* differs from the situation *in vivo* with respect to the biophysical restraints. The main difference is the crowded nature of the cytosol in living cells due to the presence of numerous different biological macromolecules and solutes [75]. The total intracellular protein concentration is estimated to be as high as 300-400 g/L [76]. Although most polypeptides have the capability of folding spontaneously *in vitro*, many proteins do not so *in vivo* but rely on other assisting proteins. Protein biosynthesis implies that the nascent polypeptide chain leaves the ribosome with its amino terminus first. As long as polypeptide synthesis is not completed the chain is highly susceptible to aggregation due to solvent-exposed hydrophobic residues unless it is sequestered from the intra-cellular surroundings by chaperonin proteins. In *E. coli*, the most important and best characterized chaperonins belong to the GroEL/GroES system [77-82]. The latter forms a large macromolecular assembly and allows folding of single polypeptides in a protected environment. Interestingly, the GroEL/GroES system is not specific for certain proteins. This might be puzzling since protein folding pathways are highly specific, depending on the amino acid sequence, whereas the molecular mechanism used by the chaperonin system is not. The situation became even more complex when GroEL/GroES, previously assumed to assist in protein *folding*, were found to promote protein *unfolding* [83-85]. However, although counter-intuitive, the unfolding function of the GroEL/GroES chaperone system is fundamental to the molecular mechanism of substrate-independent folding assistance. More precisely, functional analysis indicated that GroEL/GroES did not actively unfold polypeptide substrates, but rather shifted the equilibrium toward the unfolded state by the physical interaction with solvent exposed hydrophobic amino acids, a hallmark of un-/misfolded polypeptides. The latter might have become locked in intermediate states corresponding to local minima in the energy landscape of protein folding, and the ATP-driven chaperonin system provides much of the activation energy that is needed to overcome this energy barrier, thereby generating multiple chances for intermediates to find a productive folding pathway. In the absence of chaperone systems, the dead-end intermediates are trapped kinetically in unfavorable high energy states which would slow down the process of protein folding far beyond the life-span of living organisms. This raises questions as to the role of molecular chaperones for the folding of prion proteins and the disassembly of aggregates. The deleterious effect of misfolded prion proteins, however, is not restricted to the formation of high-molecular weight aggregates that eventually interfere with vital cellular functions. Beyond that, certain prion proteins apparently snap up and eventually outcompete the cellular refolding machinery, thereby propagating the prion state without being delivered to the proteasome system [81].

A.3 Structure Determination of Biological Macromolecules

Structural biology provides essentially four different biophysical methods to determine the three-dimensional (3D) structure of proteins, nucleic acids and other biomolecules:

(i) X-ray crystallography, (ii) NMR spectroscopy, (iii) cryo electron microscopy (cryo-EM), and (iv) neutron diffraction. All these methods use electromagnetic radiation or sub-atomic particles that interact with the molecular target, thus providing information about the electronic and magnetic properties which are then converted into structural restraints. Despite these similarities, the methods differ with respect to sample preparation, data acquisition and processing, structure calculation, the quality of the final model and the overall time scale.

A.3.1 Macromolecular Crystallography

The method of structure determination by means of diffraction experiments on single crystals dates back to the early decades of the 20th century. First used to analyze the crystal structure of small chemical compounds, X-ray crystallography had been successfully applied later also to proteins, nucleic acids and even large macromolecular complexes. With the invention of highly sophisticated X-ray sources (synchrotron beam lines), advanced software tools, and superior workstations, structure determination by X-ray crystallography now provides a very powerful tool for structural biologists. In principle it has become possible to solve the 3D crystal structure of any molecular entity, may it be as small as water in ice crystals or as large as complete ribosomes [86].

Single Crystals

Two steps in the process of crystal structure determination still can be considered bottle necks which indeed often prevent the straightforward structure solution: crystallization and phasing. Growing single crystals of good diffraction quality represents a major challenge in macromolecular crystallography. Single crystals consist of building blocks that are regularly oriented and repeated in all three dimensions throughout the crystal. The smallest repeating unit is termed ***unit cell*** that may consist, in the simplest case, of only one entity (e.g. ion-pair, molecule). The ultimate aim of X-ray crystallography is to determine the electron distribution around the atomic scaffold and, by this, the 3D structure of the molecules inside the unit cell. The main purpose of using single crystals for data collection is to amplify the signal of diffracted X-ray waves by orders of magnitudes due to the regular three-dimensional assembly of unit cells in the crystal lattice. In principle data collection would be equally well

possible on single, isolated molecules, and diffraction experiments do not necessarily rely on crystals. There are, however, major physical limitations to date: (i) X-ray waves deflected by a single molecule are extremely weak. In fact, they are so weak that the signal intensity is far below the detection limit, unless the exposure time reaches the age of the universe or the total X-ray energy applied virtually exceeds that of an atomic bomb. For obvious reasons this is out of range for any practical purposes. (ii) To collect diffraction data, the molecular target needs to be fixed in a defined orientation. For single molecules this is very difficult to accomplish, unless extreme experimental conditions (e.g. very low temperature and/or strong magnetic fields) are applied.

Electromagnetic Radiation and X-ray Diffraction

The quantum-mechanical dualism of particles and waves states that photons bear the physical characteristics of waves. As such, waves of a given wave-length (or frequency) are fully characterized by an amplitude (A) and a phase (φ). In mathematical terms, waves may be represented as complex exponentials.

$$E(t) = Ae^{i\varphi}$$

with

$$e^{i\varphi} = \cos\varphi + i \bullet \sin\varphi \text{ (Euler's Theorem)}$$

The interaction of electromagnetic radiation with matter follows the fundamental laws of electromagnetism and quantum mechanics. Depending on the wavelength of the radiation and the energy levels of the target, electromagnetic radiation may be either absorbed or deflected. Quantum theory implies that energy shows up as small packages (quanta) and absorption and emission occurs only if the radiation energy matches with the energy difference between the two electronic states involved (e.g. ground state and excited state). X-rays are very energetic with wavelengths in the range of interatomic distances. In fact, the energy of X-rays is orders of magnitude higher than that appropriate for absorption by (outer-shell) electrons. Therefore, X-rays are mostly scattered by electrons rather than absorbed.

The wave properties of photons show up in an oscillating electric and magnetic field associated with electromagnetic radiation. Charged particles interact with electromagnetic radiation, such that they become accelerated by the electric field, thus eventually oscillating with the same frequency as the incident waves. However, radiation-induced oscillation is much more efficient with electrons than with atomic nuclei, since the former have a very small mass and consequently a much higher charge-to-mass ratio. Oscillating electrons in turn

emit photons (i.e. electromagnetic radiation) perpendicular to the electric field of the incident wave. This process is known as scattering. Since X-ray waves are not polarized, the scattered waves travel in all directions. Waves scattered in the same direction interfere, that is the oscillating electric and magnetic fields add up to an even stronger field (constructive interference) or a much weaker field (destructive interference), depending on the phase differences.

In the case of repeating objects such as crystals, diffraction may be considered reflection of waves at a set of parallel lattice planes, and discrete diffraction signals are commonly referred to as reflections. In crystals, most scattered waves cancel each other through destructive interference unless the scattering of all objects is in phase, that is the phase shift is an integer multiple of the wave-length. This relationship is expressed with ***Bragg's law***:

$$n\lambda = 2d_{hkl} \bullet \sin\theta \Leftrightarrow \sin\theta = n\lambda/(2d_{hkl})$$

θ is the angle of incidence (the total angle of diffraction, however, is 2θ), λ is the wavelength, d_{hkl} is the distance between two parallel lattice planes of order hkl (also known as ***Miller indices***), and n is an integer. There is an important implication of Bragg's law: d_{hkl} and $\sin\theta$ are inverse proportional (at constant wavelength). Thus, waves scattered to large angles may be considered as being reflected at small-spaced lattice planes, thereby increasing the resolution of the molecular image that is calculated from the diffraction data. In crystallographic terms, the resolution limit is related to the minimum distance (d_{min}) between neighboring, parallel lattice planes that give rise to diffraction and constructive interference of scattered waves. Since the wavelength of X-rays is close to the interatomic distances of biological molecules the phase shift of deflected/scattered X-ray waves gives raise to interference and intensity modulation of the resulting signals. Thus, the scattered X-ray waves contain all the information about the relative positions of electrons throughout the unit cell. The number of *independent reflections* (N_{ref}) that can be measured depends on d_{min}, the volume of the unit cell (V) and the crystal symmetry, but not on the wave-length.

$$N_{ref} = 2\pi V/[3d_{min}^3]$$

Fourier Transforms and Electron Density Maps

The diffraction of X-ray waves gives rise to an interference pattern which contains information about both the symmetry properties of the crystal and the distribution of electrons within the unit cell. The crystallographic symmetry is reflected in a corresponding symmetry of the diffraction pattern, whereas the electron density distribution in the unit cell correlates

with the signal intensities. Scattering causes a phase shift of the diffracted waves with respect to the origin of the unit cell. Interference of diffracted X-ray waves therefore results in an intensity modulation of the diffraction signals. Notably, diffraction on crystals and other regularly repeating objects yields *discrete* data points (reflections) with different intensities. In mathematical terms, each reflection of order hkl is characterized by the amplitude and the phase shift of the diffracted wave.

$$F_{hkl} = |F_{hkl}| \exp[i\varphi_{hkl}]$$

F_{hkl} is known as the ***structure factor*** for reflection hkl, $|F_{hkl}|$ is the amplitude of the diffracted X-ray wave, and φ_{hkl} is the phase angle. The amplitude is related to the (measured) signal intensity of reflection hkl (I_{hkl}) by $I_{hkl} \sim |F_{hkl}|^2$. The structure factor F is equivalent to the total scattered wave resulting from interference of all scattered waves traveling in the same direction, which is the sum of all atomic scattering factors.

$$F = \Sigma f_i$$
$$f = f_0 \exp[-B\sin^2\theta / \lambda^2]$$

f_0, also referred to as ***atomic form factor***, is an atom-type specific scattering factor which is correlated with the number of electrons. θ is the angle of reflection/deflection, λ is the wavelength, and B is the temperature factor which measures an atom's displacement from the mean ($\bar{u}$) due to high mobility and/or anisotropic defects in the crystal.

$$B = 8\pi^2\bar{u}^2$$

Diffraction/scattering may be considered the physical equivalent of a mathematical operation called ***Fourier transform*** (FT) of an object. This is accounted for in the structure factor equation:

$$F_{hkl} = \int \rho(x,y,z) \exp[2\pi i(hx + ky + lz)]\, dV$$

In principle, the image of the diffracting molecule(s) in the unit cell can be calculated by a mathematical formalism known as (inverse) Fourier transform.

$$\rho(x,y,z) = \Sigma F_{hkl} \exp[-2\pi i(hx + ky + lz)]$$

$\rho(x,y,z)$ is the electron density at point (x,y,z), V is the unit cell volume, and F_{hkl} is the structure factor for reflection hkl. Notably, *every* structure factor F_{hkl} contains information about the positions of *all* electrons in the unit cell. To reconstruct the electron density of the molecular image, it is therefore desirable to include as many terms in the summation as possible. Notably, high-resolution data contain significantly more information than low-resolution data, and that is the reason why poor diffraction quality may severely interfere with model building and structure analysis. The structural model, however, depends much more on the phases than on the measured intensities. Fourier synthesis using mixed amplitudes and phases from completely unrelated objects yields indeed a noisy, yet unambiguous image of the object from which the phases were derived. Phases therefore contain a great deal of structural information, underlining the importance of highly reliable phase information. The phase information is not directly accessible by the diffraction experiment and needs to be determined by other (indirect) methods. This major challenge in crystallography is illustrated in a very instructive example given in Kevin Cowtan's book of Fourier transforms [87].

Solving the Crystallographic Phase Problem

An incident X-ray beam hitting a crystal is scattered at the electrons of the molecular target. The scattered X-ray waves overlay in all three dimensions and eventually give rise to an interference pattern on the detector. The very diffraction event causes a phase shift of the total diffracted waves relative to the incident beam depending on the distribution of electrons in the unit cell. Thus, in principle the position of all atoms inside the unit cell can be calculated, knowing the *intensity and the phase shifts* of all interfering X-ray waves. However, whereas the signal intensity can be directly measured, the phase information is *not* directly accessible. In simple terms, losing the phase information is equivalent to losing information about the precise localization of electrons and atoms inside the unit cell.

The phase problem is usually solved by modulation of the diffraction data such that the information derived from the signal intensity modulation may be used to extract initial phases. For *de novo* structure solution the most important methods are single/multiple ***isomorphous replacement*** (SIR/MIR) and single/multiple ***anomalous dispersion*** (SAD/MAD). In the case of SIR/MIR, isomorphous crystals of heavy atom derivatives are used to modulate the diffraction pattern such that the heavy atom positions in the unit cell may be derived from the intensity differences. SAD/MAD techniques rely on atoms that may give raise to anomalous scattering. For this purposes, proteins and nucleic acids are labeled with seleno-methionine (Se-Met) and 5-iodo-uridine, respectively. In the case of anomalous scattering, the phase information may be obtained from the intensity difference between $F_{(hkl)}$ and $F_{-(hkl)}$.

If the structure to be determined is similar to an already known structure, the latter may be used as a search model. This method is known as ***molecular replacement*** which provides a fast-forward approach for phasing and structure solution. Molecular replacement is routinely used in crystallography for ligand screening. In a very instructive example provided by Kevin Cowtan, the basic principle of molecular replacement is nicely depicted [87]. Starting with the diffraction pattern (FT) of an object (e.g. a cat), the task is to reconstruct the image from the diffraction data. Fourier synthesis requires knowledge about both amplitudes *and* phases. The former are related to the reflection intensities ($I_{hkl} \sim |F_{hkl}|^2$), but the phases are not directly accessible in real X-ray diffraction experiments. However, the phase problem can be solved by using a slightly different object (e.g. tailless cat) as a search model for which it is possible to calculate the amplitudes and phases of its FT. By combining the measured intensities from the original object (cat) with the phase information from the search model (tailless cat), it is possible to restore the image that has been used in this *in silico* diffraction experiment. Most importantly, this method restores the *complete* object (cat *with tail*) despite the fact that the phases from search model (*tailless* cat) have been used for Fourier synthesis. The corresponding electron density map is obtained with $|F_o|$ amplitudes. If the model phases lack structural information about molecular details of the original object (e.g. the cat's tail) the latter show up at about half of the original density. To get the full density back, an extra amplitude correction, equivalent to the difference between the observed amplitudes ($|F_o|$) of the object and those from the search model ($|F_c|$) is added to the Fourier synthesis. By that, the corrected FT corresponds to a $2|F_o|–|F_c|$ electron density map. Since $2|F_o|–|F_c|$ maps may show details that were not accounted for in the molecular model, $2|F_o|–|F_c|$ Fourier coefficients can be used to generate (almost) unbiased electron density maps. As such, ***omit-maps*** may reveal those parts of the electron density which are biased toward the model-derived phases. For practical purposes, different parts of the model are omitted for map calculation. By that, the electron density at these sites relies more on the measured intensities (and the amplitudes derived thereof) than on the model phases. If the model parts in question are correct, the omit-map looks pretty much the same as if the complete model had been considered. Therefore, omit-maps can be used to identify and/or validate the binding sites of water, ions, and small-molecule ligands. *Composite omit-maps* comprising a complete set of omit-maps, each corresponding to a small part of the model, can be used to cover the entire model. This method may be regarded as a type of ***real space cross-validation***, comparable to NOE cross-validation in NMR spectroscopy (see Appendix A.3.4, "Cross Validation"). In addition, omit-maps may be combined with a simulated annealing (SA) step prior to the map calculation to prevent model bias due to the phase information from previous refinement cycles.

Structure Calculation and Model Refinement: New Perspectives

In a recent correspondence to "nature", N. Furnham, T. Blundell, M. DePristo and T. Terwilliger brought about some general considerations regarding presentation, interpretation and the use of molecular structures studied by X-ray crystallography [88]. The main objection to the standard single-structure representation usually provided as the result of crystal structure analysis relates to both the uncertainties due to (experimental) errors during data collection, intensity read-out and phasing, and the true nature of conformational heterogeneities within the crystal lattice, all of which are difficult to assess. Consequently there is a lack of model precision (reflected in the crystallographic *R*-value) and/or accuracy, the latter describing the differences between the *calculated model* and the *real structure*. In high-resolution crystal structures, alternative conformations are modeled to account for the structural heterogeneity in the crystal. Likewise, firmly bound water molecules and ligands are taken into account. On the contrary, low-resolution structures are usually under-restrained, excluding the identification of alternative conformations and the sites where solvent molecules might have been bound. Consequently, low-resolution structures are published basically as single conformer models. However, although high-resolution models usually contain structural features deviating from the standard single conformation, thereby suggesting high structural heterogeneity, they are in fact much better restrained and thus have a (much) higher precision. Vice versa, models of low-resolution structures usually lack alternative conformations, thereby suggesting high precision and reliability, although the uncertainty of atomic coordinates is in fact (much) higher because of the weak restraining power of the diffraction data.

Terwilliger and co-authors therefore proposed to deposit crystallographic models in the form of an ensemble of structures rather than a single conformation to account for the inappropriateness of the single species representation which might be potentially misleading. The structural ensemble typically comprises a number of conformations that all satisfy the experimental restraints derived from the diffraction data. As such, the structural deviations in the ensemble may reflect the experimental uncertainty associated with data acquisition and model building. Alternatively, the ensemble may represent the structural heterogeneity and dynamics associated with the molecules in solution or in the crystal. Intriguingly, the very same ensemble representation and its structural interpretation are generally used for NMR structures.

A.3.2 NMR Spectroscopy

Physical Background

NMR spectroscopy uses the magnetic properties associated with the nuclear spin of protons (^{1}H) and other atomic nuclei with half-numbered spin quantum numbers (^{13}C, ^{15}N, ^{19}F, ^{31}P, and others). If a magnetic field is applied, the nuclear spins behave like magnetic dipoles which are oriented in the external field. The latter causes the energy level to split into a favorable low-energy and an unfavorable high-energy level, and irradiating the sample may force the nuclear spins to absorb energy and to populate the high-energy level. The radio frequency at which spin transition occurs depends on the electronic environment of the nuclei which is governed by the three-dimensional structure of the target. Data acquisition comprises a set of NMR experiments using specific pulse sequences which allow the identification and assignment of coupled nuclei. Thus, knowing the resonance frequency of each type of proton and other nuclei, it is possible to identify the pairs of coupled nuclei in NMR spectra and to derive restraints for the calculation of structural models.

Sequential Assignment and NOE Distance Restraints

The resonance frequencies of the different types of protons and other atomic nuclei are obtained by a procedure called (sequential) assignment. To perform structure calculations of a molecular target, additional medium- and long-range distance restraints from ***Nuclear Overhauser Enhancement*** (***NOE)*** data are required. Pairs of protons that are close together in space (< 6 Å) may give rise to NOE signals due to dipolar coupling, with the intensity of the NOE signal being inverse proportional to the 6^{th} power of the distance. Classification of NOE signals into different distance classes then provides structural restraints. The intensity of NOE signals, however, may become reduced due a process referred to as relaxation. Thus, strong NOE signals indicate strong coupling and short distances whereas weak NOE signals may be the result of either weak coupling and/or long proton-proton distances, or fast relaxation through magnetization transfer mechanisms, thereby (falsely) suggesting an increased proton-proton distance. Therefore the distance restraints derived from NOE signals are usually in the form of upper-limit restraints. By that, even weak NOE signals classified as long-range interactions may represent a rather short inter-proton distance in the final NMR ensemble.

Although in principle very precise, manual assignment of NOE signals by visual inspection of NMR spectra may be susceptible to false interpretation of NMR data and model bias. Starting with a limited number of presumably unambiguous NOE distance restraints generates intermediate NMR structures which might be used as reference points in subsequent

rounds of assignment and structure calculation. This procedure effectively prevents nonsense-assignments, but in turn may overestimate strong or unambiguous NOEs, thereby narrowing the conformational space explored during structure calculation. Manual NOE assignment therefore is often biased toward the results from previous rounds of structure calculation. In addition, manual assignment protocols usually do not include statistical NOE evaluation.

Automated NOE Assignment and Structure Calculation

In general, the NOE distance restraints derived from cross-peaks of NOESY spectra can be ambiguous if there are alternative assignment options. In fact, the number of NOEs that can be considered unambiguous may drop below 10% of all NOE cross-peaks [47]. In recent years, new algorithms and software packages such as CYANA (Combined assignment and dynamics algorithm for NMR applications [47], an improved version of CANDID (Combined automated NOE assignment and structure determination module) and DYANA (Dynamics algorithm for NMR applications) have been developed to account for the uncertainties of the assignment and structure calculation procedure. A major step forward in automated assignment protocols was the invention of probability-based cross-peak assignment and the use of weighted NOE restraints. For this purpose two novel concepts have been developed in the laboratory of Kurt Wüthrich: ***network-anchoring*** and ***restraint-combination*** [47]. Network-anchoring is based on the idea that a subset of NOE restraints suggests additional distance restraints, thereby representing a self-consistent set of NOEs. By this, NOE restraints that are consistent with a subset of other NOEs are considered highly reliable and therefore (almost) unambiguous. Restraint-combination is applied to reduce the impact of false NOE assignments on upper-distance restraints. For this purpose a number of NOEs are grouped and averaged to give a mean distance restraint that is calculated by statistical methods.

Whereas manual assignment generates restraints of digital character (i.e. an individual NOE is either right or wrong), CYANA determines the reliability of assignment by calculating the probability of correctness for each NOE. By this, the ambiguity of an individual NOE becomes a function of its probability of correctness which itself relies on all other assignments. As a result, an NOE restraint might be down-weighted for structure calculations, if its probability of correctness is low due to the presence of contradicting high-probability (i.e. unambiguous) NOEs which may form a consistent network of structural restraints. Automated assignment and structure calculation is essentially based on the statistical evaluation of cross-peak positions and intensities whereas manuals assignment relies on the visual inspection of NMR data by an experienced scientist. Therefore the two methods may identify different NOE restraints and the results of structure calculation may (slightly) differ.

Biomolecular NMR Spectroscopy: Pros and Cons

There are two major advantages of solution NMR spectroscopy. First, structure determination does not rely on crystals and does not even require extremely pure and homogeneous samples. Second, NMR spectroscopy in solution allows time-resolved experiments addressing questions related to the kinetic properties of the sample, including the dynamic interaction with co-factors, molecular substrates and binding partners. Protein structure determination by NMR spectroscopy, however, is demanding in terms of sample preparation and data acquisition. To perform the complete resonance assignment, the protein sample has to be specifically labeled with ^{13}C and/or ^{15}N isotopes. In case of larger proteins, additional labeling strategies are required, including amino-acid type specific labeling.

Despite the invention of more advanced labeling techniques and pulse programs the method of solution NMR spectroscopy is limited by the size of the molecular system under investigation. The limits of NMR spectroscopy are dictated by fundamental laws of physics. Since the line width or signal-to-noise ratio of NMR data is inversely proportional to the rotational correlation time (which itself increases with the size of the molecules in solution), NMR data of large molecules or molecular complexes become increasingly intractable for structure calculations aiming at high resolution or precision. As a result, the number of theoretical data points increases with the size of the system, whereas the number of data points actually measured declines. For practical purposes, NMR spectroscopy is therefore most applicable and competitive for molecular systems below 15 kDa in size. With the use of more advanced techniques in sample preparation and data acquisition even systems with 30-40 kDa in size may become reasonable targets for the *de novo* structure determination. In recent years assignment protocols for molecular systems of more than 100 kDa have been developed, allowing the identification of secondary structure elements and the calculation of the backbone conformation. These are impressive achievements, invaluable as contributions to the field of NMR spectroscopy without any doubt. However, solution NMR spectroscopy appears to be rather limited in terms of resolution and precision for structural studies of high-molecular compounds and large complexes. In the near future, ***solid state NMR spectroscopy*** might overcome at least some of the physical limitations associated with solution NMR spectroscopy. This special NMR method, together with cryo-EM, may become even more important for the structure determination of membrane proteins which still represents one of the major challenges in X-ray crystallography.

A.3.3 Data Redundancy and Number of Variables

Determination of molecular structures by X-ray crystallography or NMR spectroscopy or any other biophysical method requires a set of experimental data that are then used as restraints for structure calculations. For most macromolecular structures, however, the number of parameters to be determined is by far much larger than the sum of data points, unless the experimental data are of exceptionally good quality. To define an atom's position in space the knowledge of at least three coordinates (usually x, y, z, in Cartesian coordinate systems) are required. Crystal structure analysis is even more demanding and requires the determination of four parameters per atom: three space coordinates (x, y, z), and one isotropic temperature factor, *B*. Diffraction data with atomic resolution (d_{min} < 1 Å) may even allow the determination of 10 parameters per atom: three space coordinates, six tensors for the anisotropic temperature factor, and the occupancy number. The latter usually equals one but may become important when alternative conformations are involved. In practice, 10 experimental data points for each atom of the structure to be determined are hardly accessible. Therefore, structural information about bond lengths and angles as known from high-resolution X-ray structures of small (organic) compounds is used as additional restraints for structure calculation and refinement. By this, the number of parameters is significantly reduced, leading to an improved ratio of data-to-parameters. Crystal structures are considered reliable and of good quality if the ratio of data-to-parameters equals at least two but it may reach even higher values if the crystals are of outstanding diffraction quality, providing atomic resolution data.

A.3.4 Reliability of Structural Models

Accuracy, Precision and *R*-values

To judge the ***accuracy*** of molecular models, it would be necessary to compare the final result with the *real structure* present in solution or in the crystal [89]. However, information about the real structure is not accessible. Instead, the ***precision*** of molecular structures is measured in terms of the differences between the calculated *structural model* and the structural restraints which themselves are deduced from the experimental data.

In crystallography, the calculation of reliability indices (*R*-factors) has proven extremely useful to assess the quality of molecular structures. *R*-factors measure the quality of structural models in terms of the difference between the experimental data and the calculated model. F_{obs} and F_{calc} are the observed and calculated structure factors, respectively:

$$R_{cryst} = \Sigma \, ||F_{obs}| - |F_{calc}|| \, / \, \Sigma \, |F_{obs}|$$

The process of structure refinement aims at the minimization of *R*-values by fitting the structural model to the experimental data. Changes in atomic coordinates go along with changes of the phase angles which themselves are not directly accessible. Therefore, structure refinement is basically phase refinement. However, refinement may be biased by fitting the model to noisy data which would result in lower *R*-values as well. The problem of model bias and over-refinement has been solved by the invention of the **free *R*-factor (R_{free})** which is calculated independently from the standard crystallographic *R*-factor (R_{cryst}) [90]. R_{free} is calculated like R_{cryst} by using a randomly chosen subset of reflections (usually 5-10%) that are excluded from structure calculation and refinement. As long as the model is fitted to real data R_{free} drops along with R_{cryst}, whereas over-refinement may further improve R_{cryst} but causes R_{free} to remain unchanged. By this, R_{free} effectively serves as an independent control for R_{work}. The difference between R_{free} and R_{work} therefore reflects the quality of structure refinement: the smaller the difference, the more reliable the structural model.

Cross Validation

In NMR spectroscopy reliability indices are difficult to calculate. Instead, the precision of NMR ensembles is given as the root mean square deviation (RMSD) of the ensemble of the best calculated structures, normally the 15-25 lowest energy structures. Although in principle very useful to assess the quality of NMR structures, RMSD values, like R_{cryst}, rely on data already used for structure calculation and refinement. To avoid the problem of model bias and over-refinement, an independent quality control, like R_{free}, would be required. For this purpose, the concept of ***complete cross validation*** has been successfully applied to NMR structures [89]. The underlying principle is the same as with R_{cryst} and R_{free} in crystallography: A subset of data points is used as independent control for the process of assignment and structure calculation. As long as the structural information of this subset is represented by the majority of all other data, the calculated ensemble of NMR structures should match equally well with the subset of data points that has never been used in the refinement process. This is however no longer the case, if model bias and over-refinement is involved.

For practical purposes, complete cross validation is highly demanding in terms of processor capacity and CPU (Central Processing Unit) time. Whereas crystallographic diffraction data in reciprocal space all contain information about the position of all atoms within the unit cell, a single NMR data point (i.e. NOE) provides the structural information for exactly one special pair of atoms. Cross validation of a certain NOE (i.e. excluding this NOE from being used for structure calculation) therefore affects the nearby region much more than distant sites of the molecule. To assess the contribution of every single NOE precisely, the process of

complete cross validation demands for a large number of NOE subsets, each comprising only the single NOE to be validated. Consequently, this procedure would require as many structure calculation runs as the total number of NOEs used. Since this would exceed by far reasonable time scales, the number of NOE subsets (and thereby the number of structure calculation runs) has to be limited by dividing the NMR data into groups of 10-20 NOEs which are used as NOE subsets for cross validation. In fact, the selection of NOE groups is one of the most challenging steps if cross validation is to be applied.

Stereochemical Quality Assessment

There are additional analysis tools available for the independent cross-checking of molecular structures. For protein structures the ***Ramachandran plot*** is of major importance as it correlates the backbone torsion angles phi and psi which themselves are normally not used as restraints during structure refinement. The plot displays allowed and disallowed regions of phi/psi combinations, and by this visualizes unusual backbone conformations. The quality assessment of protein structures also includes the analysis of stereochemical properties like bond lengths, bond and torsion angles, the planarity of aromatic systems, side chain conformations, and others, and their comparison with standard values observed in high-resolution protein structures. Advanced applications for structure validation include features for H-bond analysis which may indicate unsatured H-bond donors and acceptors and consequently suggest (180°) side-chain flips to allow for the maximum number of possible H-bond interactions, and the evaluation of solvent-accessible surface regions.

A.4 Structure Prediction and Design of Small Proteins

Structural biology is highly demanding in terms of time investment and experimental expertise. This has spurred the development of mathematical algorithms for the *de novo* prediction of (small) proteins. At first glance this might be seen as straightforward approach with the precise knowledge of the amino acid sequence and the type and strength of all physicochemical interactions (essentially H-bonds, Coulomb interactions, and hydrophobic interactions) involved. In fact the task of *de novo* protein structure prediction consists basically of "simulating" the protein folding reaction *in silico*. However, the strength of the intramolecular interactions depends on both the local dielectric constant and solvation effects which are difficult to assess properly, making the simulation of protein folding insufficient for practical purposes at present days. Instead, a two-step (low- and high-resolution) conformational samp-

ling method has been developed to predict the structure of small single-domain proteins, approaching RMS deviations for all C^{α} atoms of 1.5 Å [91]. The first and most critical step consists of sampling backbone conformations that populate a free energy basin close to the native state. The second steps aims at the native-like packing of amino acid side-chains within the protein core, representing basically a "jigsaw puzzle" in which the best packing of side-chains is to be found. The first step may become a rather challenging task since the "near-native" conformations need to be within the radius of convergence of the energy minimum representing the "real" native state. This is achieved by restricting the degrees of freedom for side-chain conformations in a first low-resolution step using a rotamer representation, thereby effectively smoothing the energy landscape. Although very useful to search for significant minima, the low-resolution search lacks details that may allow the discrimination of local minima and, by that, the identification of "real" near-native conformations. This may ultimately result in the failure of finding high-quality (i.e. low energy) models in the second step.

Adequate conformational sampling therefore appears to be crucial for the *de novo* structure prediction. In fact it is more challenging than standard energy minimization since the simplified target energy function lacks the predictive power to reliably discriminate between local false minima and the near-native funnel. The situation becomes even more demanding if the conformations of side-chains are taken into account: *In silico,* the native-like protein core packing has only minor restraining power on the backbone conformation (and vice versa), rendering native-like interactions susceptible to structural perturbations. This is due to the relatively small energy differences even between completely different protein folds that are associated with local minima in the energy landscape. In contrast, "real" energy landscapes (if they were to be computed and depicted) would show a deep minimum representing the native state, clearly separated from other local minima. This is the key drawback of the target energy function used for *de novo* protein modeling.

In summary, protein structure prediction may be considered a valuable tool for structural biology, but still faces a number of major problems. The reliable prediction of the overall fold and some core interactions in small proteins may become reasonable in the next future. *In silico* modeling of active site conformations and/or interactions with binding partners, however, appears to be much more difficult to accomplish at present. Interestingly enough, modeling aims at both ***structure prediction*** and ***rational protein design***. The former implies finding the native (lowest energy) structure for a given sequence, whereas the latter means finding the lowest energy sequence for a given structure (and/or function). Protein structure prediction and modeling is nicely reviewed in [92,93].

A.5 Enzymes: Molecular Catalysts of Living Nature

A.5.1 The Chemical Reaction Coordinate

Chemical reactions proceed along the reaction coordinate in a way that is determined by the chemical nature of the reaction partners involved, the solvent, and the overall experimental conditions (e.g. temperature, pressure, pH). In the course of the reaction, bonds may be broken and new ones may be formed, the charge distribution may change, as the conformation and chemical composition of the reaction partners may change. Some of these intermediate states, related to local minima in the reaction profile, are stable enough to allow their chemical and physical characterization. In contrast, high energy states are extremely unstable and exist for only very short periods of time, typically in the range of 10^{-15} s. They are therefore usually referred to as ***transition states***. To break and form chemical bonds, to change the molecular charge distribution, or to force chemical compounds into unfavorable conformations requires energy, usually referred to as ***activation energy***. If this were not the case, the reactants involved would be transition-state-like, extremely unstable and therefore without any benefit for practical purposes. However, the amount of energy required to start the reaction may become very small. In fact, sometimes only a few light quanta or the thermal motion of the reaction partners may be sufficient.

The Gibbs free energy profile along the reaction coordinate is usually depicted in reaction coordinate diagrams, visualizing the progress along the reaction pathway. Typically, in a simple reversible chemical reaction the reactants with Gibbs free energy G^R are processed to the reaction products with Gibbs free energy G^P. The highest point in the free energy profile correlates with the transition state $T^{\neq}$. If the free energy levels of reactants, products and the transition state are set to G^R, G^P and $G^{\neq}$, respectively, the activation energy becomes $\Delta G^{\neq} = G^{\neq} - G^R$, whereas the overall change in Gibbs free energy becomes $\Delta G = G^P - G^R$.

The driving force of a chemical reaction at a given moment of time is determined by the chemical activities (essentially concentrations) of the reaction partners which usually change over the course of the reaction. This is accounted for in the chemical reaction quotient (Q), considering the reversible reaction $aA + bB \rightleftharpoons cC + dD$ where A and B are the reactants, C and D are the reaction products, and a, b, c and d denote the stoichiometric coefficients.

$$\Delta G = -RT \bullet \ln(K/Q)$$

with

$$K = [C]^c[D]^d/[A]^a[B]^b \text{ and } Q = [C_{(t)}]^c[D_{(t)}]^d/[A_{(t)}]^a[B_{(t)}]^b$$

R is the (ideal) gas constant, T is the absolute temperature in Kelvin, *K* is the equilibrium constant, and *Q* stands for the chemical activity ratio of the reaction partners at a certain moment of time away from the chemical equilibrium. Over the course of the reaction, *Q* approaches *K*. If $K < Q$ the reaction proceeds to the left, whereas if $K > Q$ the reaction proceeds to the right. The chemical equilibrium is reached when $K = Q$ ($K/Q = 1$), $\ln(K/Q) = 0$ and $\Delta G = 0$. In the case of high-energy (yet stable) reactants and/or a large negative change in Gibbs free energy, the back-reaction may be characterized by extremely low rate constants (e.g. $2H_2 + O_2 \rightarrow 2H_2O$). In general, if $K < 1$, the equilibrium is on the side of the reactants, whereas if $K > 1$, it is on the product side.

A.5.2 General Aspects of Enzymatic Catalysis

Although biochemical reactions in general may be energetically favorable due to the driving force associated with an overall negative change in the Gibbs free energy (ΔG), by far most chemical reactions in living nature do not proceed spontaneously but rely on molecular catalysts. In energetic terms, enzymes and other catalysts lower the activation energy, thereby accelerating the reactions by orders of magnitudes. In fact, the high activation energy reduces the rate constant to values which effectively block the uncatalyzed reaction. What are the molecular reasons for the huge rate enhancements of enzymatic reactions as observed in the biochemistry of living nature? In general, enzymes reduce the activation energy by lowering the free energy of molecular transition states. This is accomplished by direct enzyme-substrate (E·S) interactions and concomitant stabilization of high-energy intermediate states. In addition, enzymes often generate and/or provide activated atomic species (e.g. nucleophiles, electrophiles) which are mechanistically important for the reaction to proceed.

What are the implications of enzymatic catalysis in nature? Life as we know it would not exist without enzymes since most biochemical reactions in living cells would require much more time than the life-span of any organism presently on earth. Protein enzymes typically accelerate chemical reactions by a factor of 10^3 to 10^{12}. By this, enzymes may act as molecular switches by turning biochemical reactions on or off. In fact, enzymes play a central role since they allow synchronizing the rate constants of biochemical reactions with vital needs of the cell, such as energy storage and breakdown, DNA replication, cell division, and proliferation.

Enzyme-Substrate Interactions: Lock-and-Key Model and Induced Fit Mechanism

Enzymatic reactions are known to be highly specific in terms of both substrate recognition and reaction mechanism, and by this highly efficient with respect to substrate turnover. Thus, most of the problems associated with standard test tube chemistry such as extended reaction times, side-reactions and low product yields do not apply to enzymatic reactions in biochemistry. This raises questions as to the molecular basis for this observed behavior and the outstanding catalytic performance of enzymes. A number of models have been developed with the aim of providing a molecular description of enzyme-substrate interactions. Two models have been fairly successful and are now considered text-book knowledge: (i) the lock-and-key model, and (ii) the induced fit mechanism.

The ***lock-and-key model***, suggested by Emil Fischer in the 1890s, is very intuitive and starts with a simple idea: if an enzyme is to interact specifically with its substrate, it somehow needs to find and select the latter out of the many different compounds also present in a living cell (or in the test tube). For this to accomplish, the enzyme has to recognize the unique molecular properties of the substrate, basically the shape or, more scientifically, the 3D structure or conformation. In a quantitative description, the electronic properties (charge distribution) and the different types of molecular interactions (H-bonds, electrostatic and hydrophobic interactions) may be also considered. The substrate binding site therefore needs to be in a way complementary to the substrate. The situation is akin to the interaction between a lock and its related key in which the latter has to fit precisely into the former to fulfill its function. Although the lock-and-key model provides a neat description of enzyme-substrate *recognition and binding*, it turns out to be rather insufficient to explain the molecular basis of substrate *processing and turnover*. Particularly the lock-and-key model is not able to explain reasonably changes of substrate conformation upon binding to the enzyme, as revealed by numerous X-ray crystallography studies of enzyme-substrate (E·S) complexes when compared with the conformation of unbound substrate and free enzyme.

The observed changes in substrate and, in some cases, enzyme conformation have been accounted for in the ***induced fit model***. The model has been proposed by Daniel Koshland in 1958 and implies that in order to become processed the substrate needs to be bound in a conformation that allows *productive interaction* with the enzyme. The main point is that this may require the substrate to adopt a conformation different from the relaxed ground-state conformation in solution. The energy required for the concomitant substrate distortion comes from the specific interactions with the enzyme. What might be the reason for an enzyme to force the substrate into an unfavorable high-energy conformation upon binding? As pointed out above, enzymes accelerate chemical reactions by lowering the energy barrier (activation

energy) that effectively blocks the uncatalyzed reaction. In the course of the reaction, the conformation and chemical nature of the substrate change along the reaction coordinate with a concomitant change in the free energy level of the intermediate states. Since the highest free energy level corresponds to the transition state (TS), stabilization of TS may significantly reduce the height of the energy barrier. Thus, binding the substrate in a conformation that is closer to TS than to the ground state provides essentially much of the driving force for the enzymatic reaction. This is the mechanistic background of the induced fit model.

In almost all cases of enzyme-substrate (E·S) interactions following the above induced fit mechanism, the functional groups required for substrate processing and turnover are provided by the enzyme. They belong to a limited number of key residues (catalytic residues) which are positioned in or near the active site to interact with the substrate. There are, however, some rare examples of enzymatic reactions that make use of *substrate* functional groups as key components of the reaction mechanism. For example, substrate hydroxyl (OH) groups may be involved in the pK_a modulation of the enzyme's general acid/base catalytic residue as it is the case in some retaining glycosyl hydrolases. In this case, the induced fit is not restricted to structure and conformation alone but also involves mechanism and function. Therefore the situation in which the substrate itself provides some of the functional groups that are mechanistically important for catalysis to proceed is referred to here as ***functional induced fit***. Intriguingly, a functional induced fit mechanism might be involved in the peptidyl-transfer reaction during ribosomal protein biosynthesis: the model of the large ribosomal subunit from *H. marismortui* with charged tRNA substrates indicates that the 2'-OH group of A76 from the peptidyl-tRNA bound to the P-site may assist in the activation of the attacking α-amino-group from the aminoacyl-tRNA bound to the A-site [94].

There is controversial discussion about E·S interactions in solution: structural studies of intermediate states and modeling of E·S interactions suggest an induced fit mechanism, leading to structural rearrangements or at least conformational changes in the enzyme or the substrate or even both. This raises the question as to the substrate and enzyme conformation in the very moment of binding. The two opposing models under debate state that (i) substrate binding may occur only if the substrate (and/or the enzyme) is already in a conformation that allows productive binding and Michaelis complex formation, or alternatively (ii) substrate binding does not rely on a certain preformed conformation but is accomplished first by a (rather unspecific) initial binding event followed by more specific interactions in the Michaelis complex [95]. Although very difficult to resolve due to the time scales involved, the first model (preformed conformation and stereoselective binding) faces critical arguments from considerations about thermodynamics and chemical equilibrium, and by that does

probably not apply to most enzymatic reactions. Since the substrate conformation in the Michaelis complex is generally different (i.e. higher in energy) compared to the (relaxed) ground-state conformation in solution, the concentration of substrate molecules bearing productive binding conformation ([S*]) is in fact very low, and enzymatic turnover would be rather poor. Ideally, the substrate molecules in ground-state conformation ([S^0]) do not interfere with productive binding, in reality however, they might function as competitive inhibitors of the enzymatic reaction. Therefore, it is likely that productive substrate binding may occur equally well if the substrate (and/or enzyme) is not in a preformed conformation. The main difference is that in the first model the conformation in the productive Michaelis complex results from conformation-selective substrate binding, whereas in the second model the conformational changes are enforced by specific enzyme-substrate interactions in the E·S complex. For example, as in the case of saccharide-binding enzymes, the initial substrate recognition and binding may involve (unspecific) stacking interactions of aromatic binding site residues with the sugar rings. By this, the substrate may be allocated to the enzyme's binding site, becoming eventually tightly bound by specific H-bond interactions, concomitant with conformational changes (substrate distortion) to allow processive binding and turnover.

RNA Enzymes (Ribozymes)

Since the discovery of naturally occurring ***catalytic RNA molecules (ribozymes)*** in the early 1980s, it has become clear that enzymatic activity is not restricted to proteins alone [96,97]. The diversity of RNA enzymes was entirely unanticipated and is now considered a key to a deeper understanding of molecular evolution. So far nucleic acid enzymes built up by deoxyribonucleotides (DNAzymes) have not been observed in nature, yet such species have been obtained by *in vitro* molecular evolution strategies. In general, every chemical compound with a defined 3D structure and, by this, precisely positioned functional groups may execute a catalytic function. The chemical nature of catalysts, however, may be in fact as simple as an inorganic metal compound, such as iron-sulfur clusters, that can catalyze electron transfer reactions. In biochemistry, the catalytic function may be best accomplished by stable polymeric macromolecules, built up by a set of monomeric components that allow the sequence-directed formation of very diverse structures. In fact, this description applies perfectly to proteins, but also to RNA molecules. The latter adopt well defined structures, thereby positioning functional groups of the nucleobases and the 2'-OH group of the ribose entities in a way to execute an enzymatic function. DNA appears to be a far less effective catalyst, probably due to the missing 2'-OH group and the base-pairing which forces double-stranded DNA into a more or less sequence-independent helical conformation.

The discovery of catalytic ribonucleic acids (ribozymes) may be considered a milestone for the theory of molecular evolution and the origin of life on earth. With the invention of molecular *in vitro* evolution techniques such as ***SELEX (Systematic Evolution of Ligands by Exponential Enrichment)*** it has became possible to select and amplify RNA enzymes bearing astonishing catalytic properties, such as self-replication (i.e. template-directed RNA polymerization), carbon-carbon bond formation, and enantioselective discrimination between the D- and L-stereoisomers of amino acids [98,99]. Some RNA species are shown to combine both the catalytic activity and the capability of self-replication which are accounted the two most important molecular functions in the pre-biotic world. The unique properties of RNA form the basis for the ***RNA world hypothesis***: Since RNA may act as *catalytic enzyme and as hereditary molecule* capable of generating molecular offspring on its own, RNA is believed to have been once played a central role for the development of simple living systems. The idea is strongly supported by the existence of RNA molecules forming the catalytic core of ribonucleo-protein (RNP) complexes. Most intriguingly, the molecular architecture of the ribosome revealed that the peptidyl-transfer reaction does not rely on ribosomal proteins but is indeed catalyzed exclusively by ribosomal RNA (rRNA). By that, the ribosome is in fact a ribozyme [95,100].