5 Atoms, Molecules and Macromolecules

If you are ready, our journey will now begin. After entering Netland, we will first visit 'simple networks'.[1] In the next section, I discuss proteins, essential building blocks of our cellular networks. They carry out almost all tasks in the cell and provide most of the cell's shape, too. We shall begin with the enormous difficulties our proteins have to overcome to reach their final structure. In the next step, we examine their complex stability patterns using the idea of energy landscapes. Finally, we discover how weak links help them in all these processes.

5.1 Protein Folding Problems

Most proteins in our cells are active only in their native structure, which represents the minimal energy state of all possible shapes. These shapes are called conformations.[2] A random search for the absolute energy minimum (the native conformation) or in other words, protein folding, would require an astronomical number of approximating steps if all conformational states were probed. This has led to the famous paradox due to Cyrus Levinthal, stating that the Universe has

[1] The expression 'simple networks' reminds me to stress that all self-organized, natural networks are rather complex. In particular, the level of complexity may significantly increase with the number of bottom network layers below the level of the examined top network. From this standpoint, networks regarded as simple by common sense are indeed simpler. However, these 'simple' networks may have a much more complex structure than their 'complicated' counterparts.

[2] In a few cases, as in prion proteins, the shape with minimal energy cannot be reached by the protein, since an energy barrier prevents the protein from adopting this favorable conformation. However, in the specific case of the prion protein, we are very lucky to have this energy barrier. Without it, the prion proteins in our brain, which are protecting both me and the reader from brain damage at this very moment, would fold to their minimal energy state and aggregate. This extensive self-assembly would cause the death of our neurons.

not lasted long enough for the folding of a simple protein (Levinthal, 1968).[3]

The Levinthal paradox. To understand the details of the Levinthal paradox, we have to imagine a polypeptide chain which contains one hundred amino acids, bound one after the other by peptide bonds. Due to rotation around single bonds, each amino acid residue has many distinct conformations, depending on its main chain and side chain dihedral angles. Geometrical constraints restrict the number of these conformations. The reason for this is that two atoms cannot approach each other beyond a certain limit, and large overlaps are not permitted. Although this does indeed restrict the number of possible conformations per amino acid, the number is still higher than unity. Taking the very moderate estimate of two, the whole polypeptide chain must have 2^{100} possible conformations, which is an extremely large number. If we let the protein fold and sampling is random, in the worst case the protein will try every conformation during its folding before finding the single solution among the 2^{100} possibilities. If we allow 1 picosecond (10^{-12} seconds) for each transition between two conformational states, the time required for the folding process would be slightly more than 2^{18} seconds, which is close to 40 billion years. This time is more than double the currently estimated age of our Universe. Levinthal's paradox clearly shows that the sampling of the conformational states cannot be random and has to be guided (Levinthal, 1968). The details of this guidance will be revealed in the rest of the chapter.

The Levinthal paradox already shows that proteins have to solve a formidable problem in order to find their unique native structure. To get a visual idea of the task, imagine several hundred, rather different pieces of nano-LEGO. These will correspond to the building blocks of the proteins, the amino acids. Among the 20 types of native amino acids, the smallest amino acid, glycine has only one proton and a pair of electrons as a side chain, while the largest, the tryptophane has two complex ring systems. If you were asked to assemble several hundred of

[3]The absolute energy minimum here is in fact the lowest free energy. The free energy is defined in thermodynamics as $G = H - TS$, where G is the Gibbs free energy, H is the enthalpy (which is the energy gain due to the formation of chemical bonds during the folding process), T is the temperature measured in kelvins, and S is the entropy, the measure of disorder in the system. Therefore, reaching the free energy minimum during the protein folding process requires a simultaneous optimization of the energy gain by the formation of various bonds, and the minimization of the entropy decrease due to the organization of the protein structure.

these vastly different molecules in such a way that there was no space left between any two of them, you would already start to scratch your head. Suppose now you are told that additional rules exist: some of the nano-LEGO pieces are white (these are called hydrophilic amino acids, e.g., glutamic acid) and they have to stay on the surface all the time. Other nano-LEGO pieces are grey (these are called hydrophobic amino acids, e.g., leucine) and they all have be buried inside the molecule (see Fig. 5.1). Additionally, all amino acids are tied with a string in a specific order, and the string must not be cut since it represents the protein backbone, where the linking chemical bonds,[4] the peptide bonds, are very stabile and will not break. Finally, you are informed that you have less than one second to complete the job. I think by this time you would start to laugh (thereby wasting your precious second) and you would give up. Not the proteins! Most of them complete the assembly, which is called protein folding, in a fast and highly efficient way.

How do proteins fold? Protein folding is characterized by two main steps in vitro (see Fig. 5.1). In the first few milliseconds most of the secondary structure, i.e., individual α-helices and β-sheets, has already formed. In most cases folding starts with the formation of α-helices, since the participation of adjacent amino acids is required here. β-sheet formation establishes hydrogen bonds[5] between amino acids, which are far from each other in the primary sequence. For this reason, when β-sheets are formed, a greater entropy decrease occurs than during the formation of α-helices. At the end of this first step the hydrophobic segments are segregated by the surrounding water and form the hydrophobic core of an intermediate which is often called the molten globule. This process is known as hydrophobic collapse since the volume of molten globules becomes almost as small as that of the final,

[4]Chemical bonds are made of one or more pairs of electrons which interact simultaneously with the two atoms bound together. Electrons from the two interacting atoms will leave their original atomic orbitals and form a bond if their energy is lower in the molecular orbital.

[5]Hydrogen bonds, or hydrogen bridges are low energy chemical bonds, where the nucleus of a hydrogen atom, a proton, can be found simultaneously in the vicinity of two atoms. In biologically important molecules, these two atoms are usually nitrogen or oxygen, both having a free electron pair. The proton uses its wave properties to occupy the space beside both atoms, even if they are relatively far from each other so that a 'forbidden' region occurs between them. Besides electrons, the proton is the only building block in our molecules which still has this property to any appreciable extent. Quantum mechanical wave properties decrease with the weight of the particle. This is the reason why we do not have carbon bridges, oxygen bridges, etc., in the same sense, since they are too heavy to display such an effect.

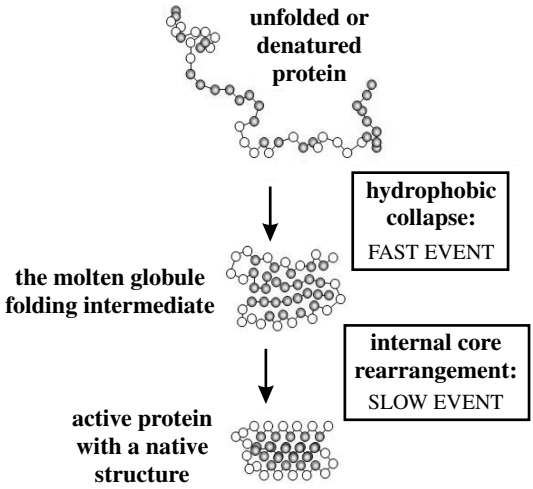

Fig. 5.1. The two major steps of in vitro protein folding on a hypothetical example of a small protein. *Grey* and *white circles* represent hydrophobic and hydrophilic amino acids of the protein, respectively. Conformational states of the target protein were adapted from Thirumalai and Guo (1994)

folded protein. If the protein is larger than 30 kDa, the molten globule is usually fairly stable. In the partially folded state of molten globules, the α-helices and β-sheets have not found their correct, tightly fitting relative position, which means that the protein does not yet have a stable tertiary structure. Molten globules still have large unburied hydrophobic surfaces. This makes them vulnerable to unspecific attachments, a process known as aggregation.

The second major step of protein folding is the slow, rate-limiting step. Here the inner, hydrophobic core of the protein is reorganized. In parallel with this, unique high-energy bonds are formed, such as disulfide bridges and ion pairs, and the peptide bonds are isomerized beside proline amino acids. The free energy gain due to these processes enables the formation of local, thermodynamically unstable, 'high-energy' protein structures, which are stabilized by thermodynamically favorable conformation of the rest of the protein. These high-energy segments of proteins can stabilize themselves by forming complexes with another molecule. They thus often serve as active centers of enzymes or as contact surfaces between the various proteins involved, e.g., in signal transduction.

Protein folding is not a straightforward process. Dead-ends, reverse reactions, futile cycles are all characteristic features of this process. A

minor amount of fully folded, native protein always coexists with various forms of molten globule and the remaining traces of unfolded protein. Aggregation of unfolded protein molecules and molten globules is a major threat that would drive the majority of folding intermediates into unproductive side reactions, long before they could reach their fully folded, competent state. Moreover, unaided protein folding often leads to folding traps. In one of these traps, the rearrangement of the hydrophobic core is often prevented. Nascent proteins have the additional problem that they have to fold before they are even ready. The first protein segment which leaves the ribosome[6] will certainly have a different energy minimum to the whole protein. Therefore, in many cases, in vivo protein folding has to be delayed, and the premature folding of the first synthesized part of the protein has to be delayed until the rest of the protein has been synthesized (Bryngelson et al., 1995; Dill et al., 1995; Dobson et al., 1994; Kim and Baldwin, 1990).

In the next two sections, I use the network approach to see how proteins are designed and assisted in accomplishing the enormous task of protein folding described above.

5.2 Energy Landscapes

In the last section we saw that protein folding is governed by a search for the conformational state of the polypeptide chain with the lowest energy. We may understand this process better if we draw all the possible conformations with their corresponding energy levels. This is called an energy landscape. Landscapes were first introduced to science in 1932 (Wright, 1932) and fifty years later the powerful concept was applied to understand protein folding (Bryngelson and Wolynes, 1987; Bryngelson et al., 1995; Dill, 1985; 1999). The landscape of protein structures is shown in Fig. 5.2. To explain it, let us go through step by step.

To make the energy landscape understandable visually, the various conformational states of the protein are viewed in two dimensions only. Thus the protein conformations occupy the (x, y) plane. The energy corresponding to each of these conformations is plotted on the vertical or z axis. The large cavity at the front represents the absolute energy minimum of all conformations. In proteins, this is usually called the

[6]The ribosome is a macromolecular complex formed by a large number of RNA and protein molecules. The ribosomes are responsible for protein synthesis in our cells.

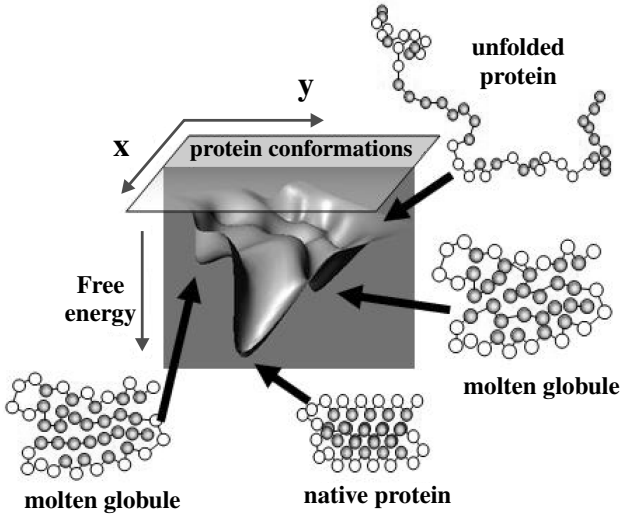

Fig. 5.2. Energy landscape. In this hypothetical energy landscape, the (x, y) plane serves to locate the various conformational states of the folding protein. The free energy of the given conformation is shown on the *vertical axis*. The approximate positions corresponding to the unfolded protein, the various molten globules, and the native protein are indicated

native conformational state of the protein. All other depressions in the landscape represent local energy minima corresponding to folding intermediaries, e.g., molten globules. The saddles between the depressions show the activation energies that the protein has to collect to get from one stable state (local energy minimum) to another. Similar landscapes can be constructed to describe the stability conditions of complex systems other than proteins, e.g., ecosystems, social networks, etc. The general term 'stability landscape' denotes all these landscapes and will be detailed in Sect. 12.2.

The energy landscape can also be visualized as a contour plot similar to those used on maps (see the top middle panel of Fig. 5.3). Here the various contour lines represent the energy levels and the actual position on the plane of the figure gives the conformation of the protein.

A network representation of the energy landscape was introduced and analyzed by Doye (2002). To illustrate this approach, the top right-hand panel of Fig. 5.3 shows the network representation of the energy landscape in the top left-hand panel. In this network, nodes

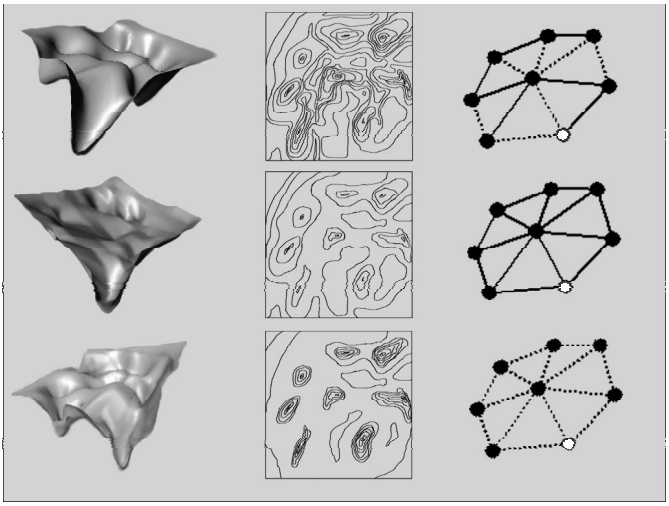

Fig. 5.3. A two-dimensional conformational space of a hypothetically simple protein is depicted to illustrate the importance of having the right amount of weak energy links in order to stabilize energy networks. The *horizontal plane* shows protein conformations and the *vertical axis* shows their energy levels. All eight energy minima remain the same in the 'normal' (*top row*), 'all-strong' (*middle row*) and 'all-weak' (*bottom row*) energy landscapes. However, the activation energies between the minima are variable, minimal and maximal in the normal, all-strong and all-weak energy landscapes, respectively. The *left-hand panels* show a 3D representation of the energy surface, whilst the *central panels* are contour plots of the energy levels around the eight minima, and the *right-hand panels* show the network representation of the transitions between the minima drawn using the concept due to Doye (2002). On the energy nets, the *white dot* represents the absolute energy minimum, while strong and weak energy links are marked with *continuous* and *dotted lines*, respectively (Csermely, 2004; 2005). Figure courtesy of Máté Szalay of the LINK group

represent energy minima and energy links[7] correspond to activation energies. The energy landscape of proteins has both a small-world and a scale-free character (Doye, 2002; Doye and Massen, 2005; Rao and Caflish, 2004; Scala et al., 2001).

Protein folding takes fractions of seconds or minutes, depending on the protein and the circumstances. The small-worldness of the energy landscape network gives us another explanation as to why our

[7]The expression 'energy link' is used to discriminate the links of the energy network – which actually denote transitions from one conformational state to another – from the links of the conventional, topological networks.

proteins fold so efficiently. Protein folding is often guided by the modularity of the energy landscape. In these cases, the landscape has a set of high activation energy barriers (weak energy links in the energy net), which exclude a major part of the possible conformational states from the folding process (Otzen and Oliveberg, 1999; Plotkin and Wolynes, 2003). Here again, we may observe a helpful mixture of global connectivity and confined relaxation. Global connectivity ensures that most starting conformations will find the single native conformation. Confined relaxation restricts further changes of the conformation once it has reached a certain level of development towards the native state.

Energy nets may provide further insights to extend our knowledge of network stability. In the energy net, an energy link can be defined as strong if the corresponding transition state in the energy landscape has low activation energy. An energy link is weak if the corresponding transition state has high activation energy. Weak energy links mean that the conformational change between the two corresponding nodes (i.e., protein conformations having a local energy minimum) has a low probability due to the high activation energy between the two energy states. If we imagine the extreme situation in which there are only strong energy links between the local energy minima, all activation energies will be low and all transitions will occur easily (middle panels of Fig. 5.3). We have an undefined (unstable) system, because the protein shifts to the global energy minimum without any appreciable transition time in local energy minima, and this is characteristic of unrestricted, fast protein folding.[8] If we have only weak energy links between the energy minima (bottom panels of Fig. 5.3), all local energy wells will behave as folding traps. The system becomes superstable, but the network itself is not defined, because we can hardly find any folding pathways in which the conformation gradually changes until it reaches the global energy minimum. The folding pathway of regular proteins corresponds to the 'normal' network, shown in the top panels of Fig. 5.3, where the distribution of energy link strength is balanced between stronger and weaker energy links (Csermely, 2004).

As my last example of energy nets I have chosen a pair of similar energy landscape networks, where the stability of two smaller systems were analyzed and compared. One of them had 19 argon atoms and the other was an assembly of 32 potassium chloride molecules (Ball et al., 1996). The argon network was grown by accommodating one or at most two atoms at a time, while the potassium chloride network was

[8] These landscapes were called buffed energy landscapes by Plotkin and Wolynes (2003).

grown by absorbing larger clumps of ions. In other words, in contrast to the argon network, the potassium chloride network was modular. An even more important difference occurred in the energy levels. The argon network showed monotonic sequences of energies, while minima of the potassium chloride network fell on multiple sequences which varied a lot, and arose from various locally stable potassium chloride crystal structures. In summary, the argon network was rather uniform, while the potassium chloride network was more diverse.

The topological differences between the networks provides a further, convincing explanation as to why potassium chloride can be condensed much more easily than argon. *"Something is not right here. Argon is a noble gas. I learned that it has very weak interactions between its atoms. As a consequence of this it does not condense well. You do not need this network hocus-pocus to explain the difference between potassium chloride and argon!"* Yes, Spite, this is true. However, argon condenses even less well than you would expect from the energy difference. This additional difference is explained by the differences between the network structures. Argon tends to stack in disordered, glass-like forms, while potassium chloride has a funnel-type energy landscape, similar to proteins, and also 'folds' (condenses) quite efficiently, like proteins and other similar macromolecules.

Weak topological links of a strong ion. I cannot resist making a comment here, in the form of two questions. What type of topological links will be more prevalent in the diverse, module-like topological network of potassium chloride than in the uniform network of argon? What may contribute to the stabilization of potassium chloride condensates? I leave the answers to you.

5.3 Weak Bonds in Protein and RNA Folding

What helps proteins to achieve their final structure so efficiently? The network approach has already given clues to find the answer. In the last section, we learned that the energy landscape of proteins simultaneously provides a confined relaxation and constitutes a small world, where the absolute energy minimum can be reached easily from any local energy minima (Doye, 2002). Small-worldness is a typical feature not only of the energy, but also of the topological networks of both globular and fibrous proteins (Bagler and Sinha, 2005; Greene and Higman, 2003; Scala et al., 2001). Interestingly, key amino acids (nucleation centers), which were shown to govern the folding process,

form highly connected hubs in this small-world topological network (Vendruscolo et al., 2002). Moreover, the small-world-type connectivity increases further during the folding process (Dokholyan et al., 2002).

In contrast to many self-organized networks, the degree distribution of protein topological networks seems to be Poissonian and not scale-free (Bagler and Sinha, 2005). Protein topological networks often have modules, which we call domains. Domains usually fold separately, have a function, and are conserved during evolution. The distribution of the folds of various domains follows a scale-free pattern (Koonin et al., 2002) implying that there is a small number of very 'popular', stable folds and that we have a relatively large number of unique, orphan folds. This is not a miracle; it is selection. Evolutionary selection preferred those structures which were both stable and folded easily. These structures are the ones which have the common feature of small-worldness.

If domains behave like network modules, it is not surprising that interdomain contacts are weak. This helps to stabilize the protein structure and gives an advantage to the complex formation of different proteins, characteristic of higher organisms. Weak interdomain contacts are often provided by fluctuating water molecules (Csermely, 2001a). Water-mediated interactions play an important role not only in binding interfaces, but also in the folding of monomer proteins. Moreover, most water-mediated interactions are long-range interactions, implying that an organized set of water molecules mediates long-range interactions between proteins that are approaching one another (Kovacs et al., 2005; Levy and Onuchic, 2004; Liu et al., 2005; Papoian et al., 2004; Pertsemlidis et al., 1999).

Molecular 'washing machines' help protein folding. In the next section, I introduce molecular chaperones. Chaperones are special proteins which help protein folding. Here I only mention them to note an important element of the mechanism whereby they fullfil their task. Indeed, the key word is water. At least one class of chaperones works like a washing machine. Using the energy gained from ATP hydrolysis,[9] they periodically stretch the misfolded protein (Chan and Dill, 1996) and then release it again to let it collapse to a more compact state. During this cyclic motion, they wash the internal, hydrophobic core of the protein through with water, again

[9]The energy of ATP hydrolysis most probably does not cause the stretch, but– like a Maxwell demon – just organizes the Brownian motion of water molecules into a 'macroscopic' motion by preventing their contribution in any other directions than the desired one.

and again, thereby accelerating its folding (Csermely, 1999; Kovacs et al., 2005).

An important mechanism whereby water helps protein folding and complex formation occurs when it acts as a lubricant, assisting protein motions (Barron et al., 1997). Although several proteins can withstand a transfer to non-aqueous media, most enzymatic functions are stopped in the complete absence of water. Lyophilized food and dried meat last longer not only because the bacteria which have invaded them are then doomed or go into a non-metabolic spore state, but also because most digesting enzymes are inhibited. Moreover, dry proteins have a 'memory'. They preserve enzyme activity, if their structure has been previously stabilized. Dry proteins 'remember' the active state because their conformational changes are frozen in the absence of water (Klibanov, 1995).

Can proteins move without water? There are quite a few, sometimes contradictory observations concerning residual protein mobility in the absence of water. On the one hand, a 'monolayer' of water molecules is needed on the protein surface to restore the dynamics of biomolecules. The dynamics emerges when individual water molecules establish the percolation of their hydrogen-bonded network (Oleinikova et al., 2005; Rupley and Careri, 1991). On the other hand, in many enzymes, a residual enzyme activity can still be observed at very low hydration levels (Kurkal et al., 2005). Detailed investigations were able to discriminate protein movements, called slaved processes, which require the contribution of water as solvent, and movements which are independent of the solvent, called nonslaved processes (Fenimore et al., 2002).

In summary, water helps protein mobility. But how does it affect protein folding? A recent paper by Peter Wolynes and coworkers (Papoian et al., 2004) showed nicely that water efficiently lowers the saddles (activation energies) of the energy landscapes and makes previously forbidden conformational transitions possible.[10,11] Water molecules form

[10]The other 'weakly linked protein folding assistants', molecular chaperones, have also been proposed to smooth the energy landscape of folding proteins by Ulrich Hartl and coworkers (Brinker et al, 2001).

[11]In agreement with the in vivo situation, in silico protein folding is also helped by lowering the saddles of the energy landscape, as has been shown in the protein models of Zhang et al. (2002).

weak links with each other and with all atoms of proteins they interact with. Therefore, we can formulate the above statement in the following way: *weak links lower the saddles (activation energies) of energy landscapes* (Kovacs et al., 2005).[12] This statement can be generalized to all stability landscapes, as I will show in Sect. 12.2.

The effect of water accords with the functional definition of weak links. Let me repeat here the functional definition of weak links from Sect. 4.2: a link is defined as weak when its addition or removal does not change the mean value of a target measure in a statistically discernible way. When he gave this definition, Berlow (1999) also stated that weak links would decrease the variation and noise in the system. Individual water molecules do not change the minima of the energy landscape. They cause no change in the stable protein structure either. However, by the addition of water molecules, several previously inaccessible conformational transitions will become much easier, and as a consequence a larger convergence to the most stable protein conformation will be possible. The final assembly of available protein conformations will not have as many trapped conformers as before and will be much more uniform, with decreased variation and noise. This process gives a rather good match with the functional definition of weak links.

I hope I am not over-explaining this phenomenon. If so, I have an excuse: this statement (weak links make the energy landscape smoother) is an extremely important statement. This will explain why weak links decrease diversity, and how can they regulate the 'punctuatedness' of the punctuated equilibrium (Gould and Eldredge, 1993; Kovacs et al., 2005). I will return to these thoughts in Sect.6.3 and give a final synthesis in Sect. 12.4.

But let us return to water for the time being. Water helps the conformational transitions of proteins. What happens if these transitions are not smooth? An avalanche will occur. Early work by Ansari et al. (1985) already showed the existence of protein quakes, i.e., the cascading relaxation of myoglobin after the photodissociation of carbon monoxide. The protein quake implies the same scale-free statistics for

[12] The content of Fig. 5.3 is in an apparent contradiction with this statement. However, the similar words conceal a great difference. In the energy net representation, weak energy links were those saddles (activation energies) which were high between two energy minima. In the topological network of protein-related physical interactions, water-mediated weak links actually transform the weak energy links of the former energy net into strong energy links, implying an easy transition between two energy minima.

protein dynamics as we have seen for other self-organized systems in Sect. 3.3.[13]

A number of protein kinetics, including the above-mentioned carbon monoxide dissociation, enzyme actions, exchange of protein protons with those of water and protein folding, are all similar to Levy flights and obey scale-free statistics (Dewey and Bann, 1992; Flomenbom et al., 2005; Metzler et al., 1998).[14,15] The scale-free statistics reflects the scale-free and hierarchical structure of the energy landscape (Yang et al., 2003), as will be described in Chap. 12. These findings mean that most protein movements are restricted and rather small. However, proteins might also make a big jump. Fortunately this only happens very rarely. This scale-free kinetics is strongly related to the scale-free structure of protein surfaces (Goetze and Brickmann, 1992) and to the presumed scale-free transport 'channels' inside the proteins (West et al., 2002).

Proteins are stabilized by weak bonds. Both secondary and tertiary protein structures are stabilized by weak, rather dispersive forces: hydrophobic bonds, hydrogen bonds and van der Waals forces. Strong bonds, such as disulfide bridges or salt bridges, may exert more influence but are used rather sparingly to achieve structural stability. Why are strong bonds missing? The network approach may explain this apparent discrepancy. A greater involvement of strong bonds would induce a lower cooperativity in the folding process and would make the isolation of the local energy minima in the energy landscape much stronger.

Ribonucleic acids (RNAs) are stabilized by weak bonds. Similarly to proteins, RNA structure can also be perceived as a mixture of strong and weak bonds, where strong bonds are represented by the hydrogen bonds between base pairs and weak bonds are the van der Waals forces between various segments of the RNA molecule. It is the latter which gives the

[13]It would be interesting to check whether protein quakes are really bigger in the absence of water than in its presence, as expected.

[14]The scale-free temporal pattern resembles that of pink noise, which was highly correlated and had a 'memory' (see Sect. 3.1). In agreement with this, a memory landscape of single enzyme molecules has been suggested (Lu et al., 1998; Edman and Rigler, 2000).

[15]It should be noted that the scale-free distribution becomes more complex in most cases, since the underlying protein structure is hierarchical and modular.

3D RNA structure its unique shape and stability (Csermely, 1997; Sclavi et al., 1998).

Protein complexes are also stabilized by weak links. Due to recent advances in the detection limits of experimental techniques, we have more and more examples showing that protein–protein (and most probably, protein–RNA, protein–DNA) complexes are indeed stabilized by weak links between the complex-forming molecules (Smith et al., 2004; Swanson et al., 2003).

We have reached the end of the first trip into Netland. It is time to have a look at the souvenirs among our luggage. What have we learnt? Simple, chemical systems like condensed potassium chloride molecules or larger entities like proteins gave the first examples of the stabilizing role of weak links. Weak links were forces or chemical bonds in these cases. Moreover, weak links were shown to be efficient at smoothing the energy landscape and therefore facilitating the transitions of the whole network from one state to another. Our next journey will take us one level higher. Proteins, which were the networks in this chapter, will form the elements of the networks in the next. What is the network then? The first system which reached the complexity we call life: the cell.