

# Chapter 1

## Introduction

Proteins are an important class of biomolecules. They are encoded in genes and expressed in cells via genetic translation. Proteins are life-supporting (or sometimes, destructive) ingredients and are indispensable for almost all biological processes. In order to understand the diverse biological functions of proteins, knowledge of the three-dimensional (3D) structures of proteins and their dynamic behaviors is essential. Unfortunately, these properties are difficult to be determined either experimentally or theoretically. The goal of computational structural biology is to provide an alternative, or sometimes, complementary, approach to protein structures and dynamics by using computer modeling and simulation.

### 1.1. Protein Structure

A protein consists of a sequence of amino acids, typically several hundreds in length. There are 20 different amino acids. Therefore, millions of different proteins can be formed with different amino acid sequences and often different functions. In the human body alone, there are at least several hundreds of thousands of different proteins, with functions ranging from transporting chemicals to passing electrical signals, from activating cellular processes to preventing foreign

intrusion, and from forming all kinds of molecular complexes to supporting various physical structures of life.

### 1.1.1. DNA, RNA, and protein

Not all sequences of amino acids are biologically meaningful. Those used in proteins are selected or decided by the biological systems. They are encoded as genes in the DNA sequences and expressed in the cells at certain times and places. A typical gene expression process occurs as follows: a gene, as a DNA sequence, that encodes a protein is first transcribed into a corresponding RNA sequence; the RNA sequence is then translated into an amino acid sequence required by the protein (Fig. 1.1).

A DNA sequence is made of two complementary chains of four different deoxyribonucleic acids — known as adenine (A), cytosine (C), guanine (G), and thymine (T) — with A in one of the chains pairing with T in another, and C pairing with G. The two chains wrap around each other and form a so-called double helix. During the transcription process, the double helix unwinds, and one of the strands is used as a template to make a single chain of RNA, which consists of four ribonucleic acids — A, C, G, and U (uracil) — that correspond to the deoxyribonucleic acids — A, C, G, and T, respectively — in the DNA templates. In this sense, the RNA sequence is equivalent to the DNA sequence; it is only transcribed in a different form. The RNA

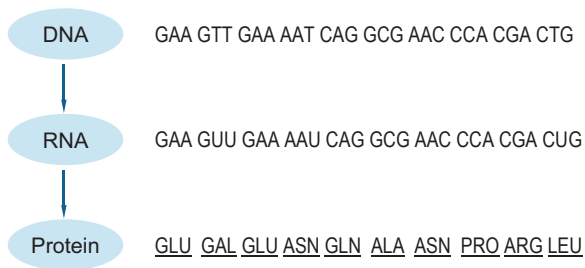


Fig. 1.1. Central dogma of molecular biology. A DNA sequence is transcribed into an RNA sequence, and an RNA sequence is translated into an amino acid sequence, which forms a polypeptide chain and folds into a protein.

		(Second Base)					
		U	C	A	G		
(First Base)	U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr End End	Cys Cys End Trp	U C A G	
	C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G	
	A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G	
	G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G	

Fig. 1.2. Genetic code. A sequence of three RNA molecules translates into one amino acid; there are 20 different amino acids that can be translated from RNA.

chain is processed to produce a chain of amino acids, with every three contiguous ribonucleic acids used to make one amino acid following a generic translation code (Fig. 1.2). There are a total of 64 different RNA triplets. They are mapped to 20 different amino acids. Two amino acids make a dipeptide. The chain of amino acids generated from the RNA chain makes a polypeptide. It folds into a unique 3D structure to become a functional protein.

Here, the DNA sequence, or the gene, determines the RNA sequence, and the RNA sequence then determines the amino acid sequence. However, the amino acid sequence, or the polypeptide, will not function as a normal protein until it folds into an appropriate 3D structure called the native structure of the protein. The latter step is called protein folding, and has been a fundamental research topic in biology because of its obvious importance in the study of proteins.

### 1.1.2. Hierarchy of structures

When forming the 3D structure, the chain of amino acids of the protein is not broken. Neighboring amino acids in the chain are always connected by strong chemical bonds. This connected chain of amino acids forms the primary structure of the protein. Different parts of the chain may form different types of structures, depending on their amino acid sequences and sometimes their interactions with other parts of the chain. The most commonly seen structures are  $\alpha$ -helices and  $\beta$ -sheets. An  $\alpha$ -helix is a helical type of structure, most often right-handed. Usually, about 3.6 amino acids form one circular section of the helix, with a 5.4 Å elevation on average. A  $\beta$ -sheet is a pleated sheet type of structure formed by a group of chain sections stretched and aligned in parallel. These structures are called the secondary structures of the protein. The secondary structures assemble themselves to finally form the overall structure called the tertiary structure of the protein (Fig. 1.3).

The primary structure shows the connectivity of the amino acids in sequence. It contains all of the information that defines the protein and hence its structure. At this level, the protein appears as a linear

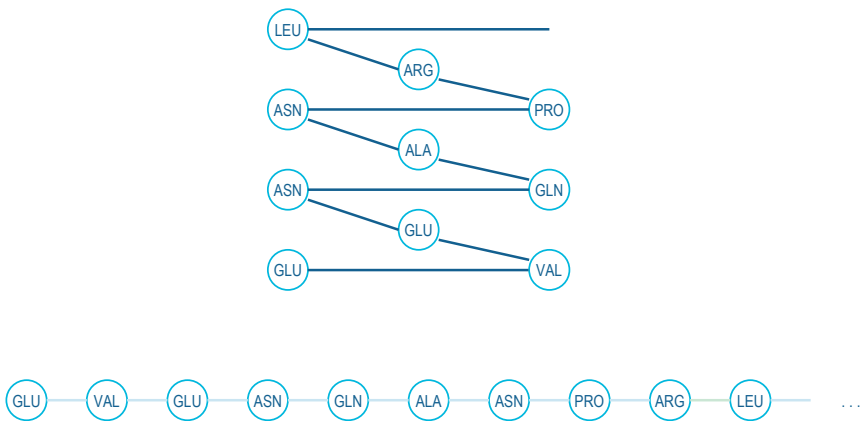


Fig. 1.3. Protein folding. A polypeptide chain becomes a functional protein after it folds into a proper 3D structure; the latter is called the native structure of the protein.

chain of amino acids. The secondary structure demonstrates the protein's local structural patterns. The protein can then be viewed at a higher level as a set of  $\alpha$ -helices and  $\beta$ -sheets pieced together. The segments between secondary structures usually do not have regular forms; they are called loops or random coils. The tertiary structure gives the detailed arrangement of all the segments of the protein in 3D space. It is usually a compact structure, and is unique for each protein. Most importantly, it is physically or biologically the most favorable structure that the protein takes and is a structure necessary for the protein to achieve its physiological function.

### 1.1.3. *Properties of amino acids*

Amino acids are the building blocks of proteins. The properties of the amino acids enable proteins to form certain types of secondary and tertiary structures and to have various biological and chemical functions.

There are 20 different amino acids. They all have an amino group ( $\text{H}_2\text{N}$ ), a carboxyl group ( $\text{COOH}$ ), and a side chain (R). A carbon atom called  $\text{C}_\alpha$  connects these three parts. The side chains are different for different amino acids. Except for glycine whose side chain has only a hydrogen atom, all side chains have a carbon atom called  $\text{C}_\beta$  connected to  $\text{C}_\alpha$ .

When two amino acids are connected to form a dipeptide, the carboxyl group of one amino acid interacts with the amino group of another and makes a covalent bond called the peptide bond. Each amino acid part between two peptide bonds in a polypeptide chain is also called a residue. Since the connection happens only between the amino group and the carboxyl group of the neighboring amino acids, the amino acid chain without the side chains is called the main chain or the backbone of the protein.

Within each amino acid residue, some local structures are relatively easy to be determined and they do not change. For example, the bond lengths and bond angles can be determined based on the chemical knowledge of the bonds and are relatively fixed. Therefore, the amino group and the carboxyl group, as well as some parts of the

side chains, can be determined and considered as rigid. However, the bonds connecting  $C_{\alpha}$  to the functional groups and some of the bonds in the side chains are flexible and can rotate. The angles around the flexible bonds are called the dihedral angles or the torsion angles. They are the main structural freedoms to be determined for the formation of correct protein folds.

Amino acids are not just the building material of proteins, but are also the driving force for proteins to form different levels of structures. For example, the hydrogen bonds between amino acids in the neighboring coils of  $\alpha$ -helices or in the neighboring strips of  $\beta$ -sheets are the key factors for proteins to form these secondary structures. Also, the fact that amino acids can be either hydrophobic or hydrophilic and proteins tend to form tertiary structures with hydrophobic cores has been considered as one of the guiding principles of protein folding.

Amino acids can be grouped according to the hydrophobic or hydrophilic properties of their side chains. One group includes the amino acids with nonpolar side chains, which make the amino acids hydrophobic. The hydrophobic group contains glycine, alanine, valine, leucine, isoleucine, methionine, phenylalanine, tryptophan, and proline. Another group includes the amino acids with polar side chains, which make the amino acids hydrophilic. The hydrophilic group includes serine, threonine, cysteine, tyrosine, asparagine, and glutamine. Acidic amino acids are those with side chains that are generally negative in charge because of the presence of a carboxyl group, which is usually dissociated (ionized) in cellular pH. They include aspartic acid and glutamic acid. Basic amino acids have amino groups in their side chains that are generally positive in charge. They include lysine, arginine, and histidine. These last two types of amino acids may be considered as electrically charged amino acids, and are hydrophilic.

#### 1.1.4. *Sequence, structure, and function*

There are close connections among sequences, structures, and functions. Sequences determine structures, and structures determine

functions. In a broader sense, this implies that similar sequences often have similar structures, and that similar structures often function similarly, although conversely it is not always so. From the point of view of evolution, genes mutate as biological systems develop. Genes that belong to close families are certainly similar and the proteins expressed from them should carry similar structures as well. The conserved parts of the structures are very likely to correspond to some biological functions shared by the similar genes and hence the proteins. These relationships among sequences, structures, and functions are fundamental questions investigated in modern molecular genetics. They are important properties that are often employed in structural and functional studies.

Here, we look at two example proteins — a retrotranscriptase found in the HIV virus (Fig. 1.4) and a prion protein that causes mad cow disease (Fig. 1.5) — and see what their sequences, structures, and functions are and how they are related.

The HIV virus is a retrovirus, meaning that it is composed of an RNA sequence. Once the virus invades a cell, the RNA sequence is transcribed back to a DNA sequence, which then integrates with the



Fig. 1.4. HIV retrotranscriptase. The enzyme transcribes virus RNA back to DNA so that the virus can be integrated in the host genome; it is 554 residues long and has 4200 atoms without counting the hydrogen atoms.

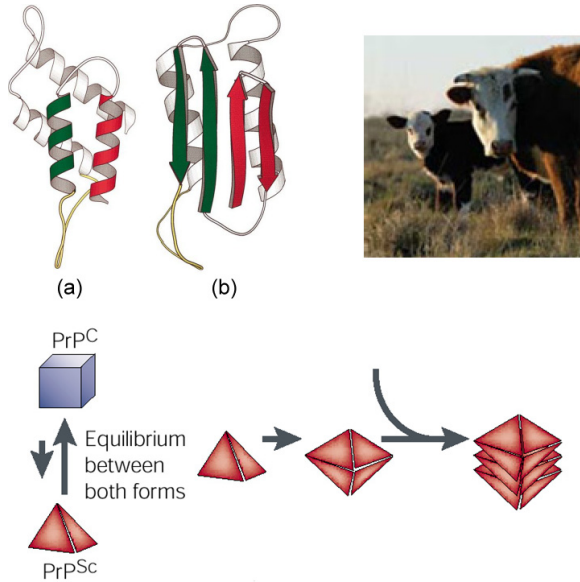


Fig. 1.5. Prion transformation. A normal prion ( $\text{PrP}^{\text{C}}$ ) can be transformed into an abnormal form ( $\text{PrP}^{\text{Sc}}$ ). The latter may be aggregated to dysfunctional complexes that can damage neuron cells and cause mad cow disease.

cell's normal genome so that the viral RNA can be reproduced. Here, the RNA-to-DNA transcription is called retrotranscription, and it occurs only in the presence of a protein called the retrotranscriptase. It turns out that the viral RNA itself contains a gene that produces the required retrotranscriptase. Retrotranscriptase is generated when this gene is expressed in the cell. The former binds onto the viral RNA and transcribes it into a DNA sequence and so on and so forth. The transcriptase protein has a special 3D structure so that it can bind properly onto the virus. The function of the protein therefore depends heavily on its 3D structure. On the other hand, knowing this dependency, researchers have been able to develop drugs that can destroy the 3D structure of the protein and obstruct its proper function, thereby preventing the virus from being transcribed. This is an example of how the structure of a protein is related to its function, and how it can sometimes be used for medical purposes.

The prion protein has been studied extensively, not only because it is related to the cause of mad cow disease, but also because it gives rise to a rare case of protein infection rather than the conventional viral or bacterial infection. The prion gene is short and expresses a small protein with around 200 residues. The normal prion protein exists in regular cells. Its function is not well understood. It folds into a simple structure with two  $\alpha$ -helices and three  $\beta$ -sheets. However, the protein sometimes misfolds into a structure with many  $\beta$ -sheets. The latter may affect a normal prion, causing it to also take on the same misfolded shape. If the process continues, a large number of misfolded prions accumulate and can damage the cells, in particular the neurons in the brain, causing mad cow disease. This is an example of how structure can affect a protein's function when it is not folded properly, and again how important this can be in medical research.

## 1.2. Structure Determination

With the completion of the genomic sequencing of human and many other species, studies on proteins, the end products of gene expression, have become urgently more important for the interpretation of genes and their implications on life. However, to understand proteins and their functions, it is essential to know their 3D structures, which, due to various technical reasons, are generally difficult to be determined. This has therefore created a research bottleneck yet to be overcome.

### 1.2.1. *Experimental approaches*

There is no direct physical means to observe the structure of a protein at a desired resolution, for example, at the residue level. Several experimental approaches have been used to obtain some indirect structural data upon which the structures may be deduced. For example, the diffraction data for a protein crystal can be obtained by X-ray crystallography, and can be used to find the electron density distribution and hence the structure of the protein; in addition, the magnetic resonance spectra of the nuclear spins in a protein can be detected by nuclear magnetic resonance (NMR) experiments and can be used to estimate

the distances between certain pairs of atoms, and subsequently the coordinates of the atoms in the protein. In both cases, computation plays an important role in collecting and analyzing data and in forming the final structures.

X-ray crystallography and NMR spectroscopy are major experimental techniques used for structure determination. Surveys on the protein structures deposited in the Protein Data Bank (PDB) show that 80% of the structures were determined by X-ray crystallography, 15% by NMR, and 5% by other approaches. These structures, about 30 000 in total, contain a high percentage of replications (structures for the same protein determined with different techniques or under different conditions). Some structures are also very similar because there are only few mutations among them. Without counting the replications and the genetically highly related structures, there may be only around several thousands of different proteins whose structures have been determined. However, there are at least several hundreds of thousands of different proteins in the human body alone. Most of their structures are still unknown.

The experimental approaches have various limitations. For example, X-ray crystallography requires crystallizing the protein, which is time-consuming and often fails. To obtain accurate enough signals, NMR experiments can only be carried out for small proteins with less than a few hundred residues. Therefore, the number of structures that can be determined by these experimental approaches is far from adequate with respect to the increasing demands for structural information on the hundreds of thousands of proteins of biological and medical importance.

### 1.2.2. *Theoretical approaches*

Theoretical approaches have been actively pursued as complementary or alternative solutions for structure determination. They are based on physical principles and therefore also called *ab initio* approaches. For example, with molecular dynamics simulation, one can compute protein motion by solving a system of equations of motion in the force field of the protein. Then, ideally, if the entire motion of protein

folding can be simulated, the structure that the protein eventually folds into may be obtained when the simulation reaches the equilibrium state of the protein. Unfortunately, folding is a relatively long process. The simulation requires a sequence of calculations which is so lengthy that the simulation cannot be completed in a reasonable amount of time, using current simulation algorithms and available computers.

An alternative approach is to disregard the physical process of folding and directly search for the protein's native fold in the protein's conformational space by using an optimization algorithm. This approach assumes that the structure into which the protein folds has the least potential energy. Therefore, a structure that minimizes the potential energy of the protein is sought. The native fold of the protein should correspond to the global minimum of the potential energy. Here, similar to molecular dynamics simulation, where the force field must be given for the protein, a potential energy function needs to be defined so that the potential energy of the protein can be calculated (and optimized) accurately. However, even if such a function was available, it could still be difficult to find the global minimum of the function.

Work has been done and progress has been made in the development of efficient algorithms so that the simulation of protein folding and the search for the global minimum of protein potential energy may eventually become feasible. For example, algorithms have been developed to increase the step size so that large-scale motions can be simulated in a given time. Furthermore, information on the end structure has been used to guide the simulation. Parallel computation has been employed to accelerate the simulation. Additionally, reduced protein models have been introduced so that dynamics simulation and global optimization may be performed more effectively and efficiently at a level higher than the atomic level.

### 1.2.3. *Knowledge-based methods*

There is another group of active approaches for structure determination that can be classified as theoretical, but is instead grouped

separately as knowledge-based because they are based on the knowledge of known protein structures rather than physical theories. In these approaches, proteins are analyzed and compared with proteins of known structures at either the sequential or structural level. Their structures can then be modeled by using the known protein structures as templates. Because the proteins need to be compared with the existing ones, these approaches have also been called comparative approaches.

The theory behind the comparative approaches is that genetically closely related proteins should have similar functions and structures. Therefore, if two proteins have similar sequences of amino acids, their structures should look alike, and one may be used as a template for another. The approaches that rely completely on genetic similarities among proteins are also called homology modeling, because the genetically closely related proteins are called homologs. Sequence comparison is not hard in general, although the sequences need to be properly aligned and appropriate scoring functions have to be developed for the evaluation of the similarity between two given sequences.

Not all similar proteins correspond to similar sequences. In fact, there are many cases where different sequences lead to similar proteins structurally and functionally. In order to find such similarities among proteins and utilize them for structural modeling, more information is required rather than just sequences. Various techniques including structural alignment have been developed to compare proteins in terms of both sequential and structural similarities. They have been applied to inverse folding as well, i.e. to find all sequences or proteins that can fold into a given structure (as opposed to finding the sequences or proteins of known structures with which a given sequence or protein may share a similar fold).

#### 1.2.4. *Structural refinement*

Due to experimental errors, the structures determined by X-ray crystallography or NMR spectroscopy are often not as accurate as desired. Further refinement of the structures, including human intervention, is

always required. For example, based on the electron density distribution generated by X-ray crystallography, the positions for the atoms in the protein can be assigned, but they are usually not very accurate. One way to improve the structure is to use an empirical energy function for the protein to adjust the positions of the atoms so that the energy of the protein can be minimized.

In NMR, in addition to obtaining additional experimental data to further improve the structure, several iterations may be required to obtain an ensemble of structures that can eventually satisfy the experimental distance data. The latter may require the solution of a nontrivial mathematical problem called the distance geometry problem. Nonetheless, the experimental data may not be sufficient for the complete determination of the structure. A subsequent energy minimization step may also be necessary. Even so, many NMR structures are still not as accurate and detailed as X-ray structures. Further justification of the structures remains an important research issue.

The refinement of comparative models presents even greater challenges. First, the structures can provide correct models for most but not all local regions of proteins, even if they are obtained from proteins of high sequence and structural similarities. Second, without further experimental evidence, it is difficult to judge whether a model is truly correct or there is still much to improve. If a model is indeed close to the true structure, refinement will be possible; otherwise, further improvement of the model will not differ much from *ab initio* structural determination. A refinement algorithm that can only provide small improvements on the model will not be of much help.

### 1.3. Dynamics Simulation

A protein structure changes dynamically, not only during folding, but also at equilibrium. The conformational changes of a protein over a certain time period are called the general dynamic properties, while the average behaviors of the protein around an equilibrium state are called the thermodynamic properties. In many cases, the dynamic and thermodynamic properties of proteins are just as important as their

structures for the study of proteins, but they are even harder to examine experimentally.

There are two ways of conducting protein dynamics simulation. One is of a more stochastic nature, using the so-called Monte Carlo method to obtain a large number of random samples of physical states. The general behaviors of the protein are then evaluated with the samples. Another is based on the simulation of the motions of the particles, namely, atoms, in the protein through the solution of a time-dependent system of equations. A phase space trajectory can be obtained from the solution, with which various dynamic properties of the protein can be analyzed.

### 1.3.1. *Potential energy and force field*

A force field can be defined based on the physical interactions among the atoms in a protein. For a given conformation, the force on each atom can then be computed as a function of the positions of all the atoms. A potential energy function can also be defined such that the negative derivatives of the function with respect to the positions of the atoms equal the corresponding forces on the atoms. The potential energy and force field must be computable one way or another in order to perform dynamics simulation.

If the forces for all the atoms are equal to zero, the atoms will stop moving. The protein is said to be in an equilibrium state. In an equilibrium state, the protein either stays in one conformation or, in most cases when the kinetic energy is not equal to zero, oscillates around the equilibrium state. Since the forces are the negative derivatives of the potential energy, at the equilibrium state, the potential energy is at a stationary point, most likely an energy minimum. At an energy minimum, the protein should be relatively stable because it has the lowest energy at least within a small neighborhood around the minimum. In general, it is assumed that the protein native conformation has the lowest potential energy in the entire conformational space or in other words is a global energy minimum, and therefore should be in the most stable state.

In principle, the potential energy for a molecule can be computed with quantum chemistry theory. However, the required calculation increases rapidly with an increasing number of atoms and electrons in the molecule. A protein may have several thousands of atoms and several magnitudes more of electrons. Therefore, it is not possible to use the quantum chemistry principle to obtain the potential energy for the protein. A general approach is to use semiempirical functions to calculate the potentials for the protein approximately. The accuracy or the quality of the functions depends on the parameters chosen for the functions. They require carefully collected experimental data to approximate.

In fact, whether or not a protein is in the most stable state should be evaluated in terms of free energy rather than simply the potential energy, because free energy describes more accurately the thermostability of a system. The free energy of a system is in general the internal energy of the system minus the product of the temperature and the entropy. So, at a fixed temperature, the free energy may not be minimized at the potential energy minimum since the entropy may increase the free energy. However, since the entropy is difficult to be evaluated, in practice, only the potential energy is considered as an approximate assessment of the stability of the protein.

### 1.3.2. Monte Carlo simulation

A Monte Carlo approach to dynamics simulation assumes that the physical state of the system to be simulated is subject to the Gibbs–Boltzmann distribution. In other words, if  $E(x)$  is the potential energy of the system and  $x$  is the state variable, then the probability of the system to stay at state  $x$  is  $p(x) = \exp[-E(x)/k_B T]/Z$ , where  $k_B$  is the Boltzmann constant,  $T$  is the temperature, and  $Z$  is the normalization constant for the distribution.

Based on this assumption, the dynamic behaviors of a physical system can be simulated by generating a large set of sample states that are consistent with the Gibbs–Boltzmann distribution of the system. The statistical properties of the system can then be easily obtained from the samples. Note that the sampled states reflect the probability

distribution of the system, but not the dynamic changes over time; hence, they are time-independent. The statistical significance certainly depends on the sufficiency of the sampling. For a large system such as a protein, the sufficiency is sometimes hard to be achieved.

The Monte Carlo simulation is temperature-specific. At a given temperature  $T$ , a random state  $x$  is generated. The state  $x$  is evaluated and taken with a probability  $p(x)$  as defined earlier. After the state is evaluated, the next state is generated and tested again, and the whole process is repeated. In this way, the set of accepted states will be subject to the Gibbs–Boltzmann distribution of the system. Because the normalization constant does not change for a given system, the probability function can be calculated without dividing  $Z$ . At each step, the next state is usually generated by making a small perturbation on the current state, simulating the change of the system from one state to another.

### 1.3.3. *Solution of equations of motion*

The major simulation scheme to be discussed in this book is based on the solution of the equations of motion that can be established for the particles in a physical system or, more specifically, the atoms in a molecule. The principle comes from classical mechanics as it is applied to the atoms in the molecule. For each atom  $i$ , assuming that Newton’s law of motion holds, the mass  $m_i$  times the acceleration  $a_i$  of the atom should be equal to the force  $f_i$ , i.e.  $m_i a_i = f_i$ ,  $i = 1, 2, \dots, n$ . Given the fact that  $f_i$  is a function of the state variables  $\{x_i\}$ , with  $x_i$  being the position vector of atom  $i$  and  $a_i$  the second derivative of  $x_i$  with respect to time,  $m_i d^2 x_i / dt^2 = f_i(x_1, x_2, \dots, x_n)$ .

The above system may have an infinite number of solutions. Only if some additional conditions are imposed, will it have a unique solution. The system may have thousands of equations for proteins. The right-hand side functions are also highly nonlinear. Therefore, solving the system of equations is not trivial. The only way to approach it is to solve it numerically. But still, the simulation can be very time-consuming or in other words computationally very demanding, because millions or even billions of steps are required to complete the

simulation of some motions of biological interest, where each step involves the calculation of the positions of all the atoms at a particular time. The reason that the simulation requires so many steps is that the time step has to be very small on the order of femtoseconds to guarantee the accuracy or convergence of the calculations, while many protein motions are on the order of milliseconds or seconds. Because of this nature, protein dynamics simulation remains a challenging research subject to ultimately become a computational tool accessible to most dynamic behaviors of proteins of biological importance.

There are two types of conditions under which dynamics simulation can be performed: initial conditions and boundary conditions. Initial conditions are usually the initial positions and velocities of the atoms, while the boundary conditions are the initial and ending positions of the atoms. The first set of conditions is used to find protein motions with a given initial conformation and temperature. A solution trajectory can be obtained to indicate how the protein conformation changes along a certain pathway under the given initial condition. The second set of conditions is used to find a possible trajectory for the protein motion between two given conformations, for example, how a protein transits from one state to another. Both sets of conditions have some important biological applications. However, the system may be easier to be solved for the first set of conditions. The system under the second set of conditions may have either multiple solutions or no solution at all. The solution method is more complex in general, but is easier to implement on parallel computers, which can be an advantage for large-scale applications.

#### 1.3.4. *Normal mode analysis*

Besides the regular trajectory calculation, an important topic in dynamics simulation is the evaluation of the structural fluctuation of a protein around its equilibrium state, which is often called normal mode analysis. A protein further changes its conformation even after reaching its equilibrium state. The reason is that although the potential energy is minimized, the protein still has kinetic energy, which

drives the system away from equilibrium, just like a simple harmonic oscillator vibrating around its equilibrium position.

The fluctuation is different along different directions in conformational space. A linearly independent set of directions can be found for the protein via normal mode analysis, each with a different vibration frequency called the normal mode. Importantly, in contrast to regular trajectory calculation, normal mode analysis can be performed analytically based on the singular value decomposition of the Hessian matrix of the potential energy function at the equilibrium state. The largest singular value corresponds to the fastest vibration frequency.

The vibration of each individual atom is a linear combination of the vibrations in all of the modes. However, if only a few slow modes are included, the fluctuations at a coarse level can then be observed without detailed fast vibrations. Such dynamic properties are often useful for the study of the most important motions of a system. In fact, the slow motions are not affected so much by the fast modes. Therefore, by using only a few slow modes, the size of the variable space for the protein is significantly reduced.

The structural fluctuations of a protein around its equilibrium state can also be obtained by averaging the structural changes obtained in regular dynamics simulation. They can also be derived from Monte Carlo simulation by averaging the fluctuations in the entire ensemble of sampled states. Additionally, there are coarse-level approximation methods such as the Gaussian network model to evaluate fluctuations at only the residue level, with an approximate elastic network model for the residue–residue interactions.

## 1.4. The Myth of Protein Folding

Having puzzled scientists for decades, the problem of protein folding remains a grand challenge for modern science. While it is a fundamental problem in biology, its solution requires knowledge beyond the traditional field of biology and has motivated research activities across many other disciplines including mathematics, computer science, physics, and chemistry.

### 1.4.1. Folding of a closed chain

First, let us consider the simple mathematical problem of folding an open or closed chain. Suppose that we have a chain with different lengths of links. We call the chain open if the two ends of the chain are not connected; otherwise, we call it closed. It is easy to fold a chain into a line (like folding a necklace into a thin box) if it is open. However, it is not as trivial to fold if the chain is closed. For example, it is harder to put a necklace in a thin box if it is still locked.

Mathematically, folding an open chain is equivalent to finding the positions for a sequence of points (connections in the chain) in a real line so that the distances between the neighboring points are equal to the given lengths (links in the chain). The problem of folding a closed chain has only one more condition that also requires the distance between the first and last points to equal zero. It turns out that this last condition makes the problem harder to solve (Fig. 1.6).

In terms of computational complexity, the problem of folding an open chain can be solved in polynomial time, while the problem of folding a closed chain has been proven to be NP-complete. In other words, an open chain can be folded efficiently in order of  $n^r$  steps,

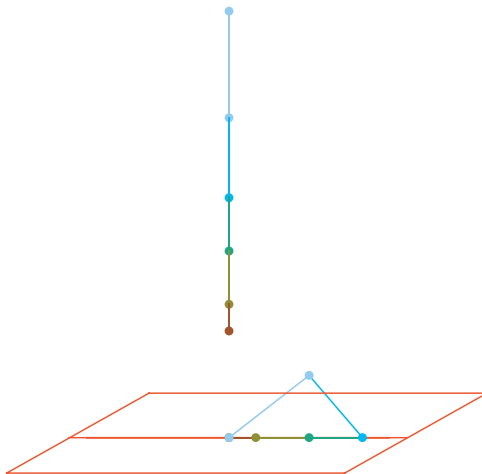


Fig. 1.6. Chain folding. The problem of folding a closed chain is NP-complete.

where  $n$  is the number of links in the chain and  $r$  is a finite number (e.g.  $r = 1$ ). On the other hand, this can be done for a closed chain only if it is on a nondeterministic computer (nondeterministic polynomial). It would run into exponentially many steps if it is on a regular (deterministic) computer, like what we have.

Folding a protein is certainly a much more complicated problem, but it indeed seems like folding a chain, only with more complex conditions. It is difficult or perhaps irrelevant to show whether protein folding is NP-complete. However, since even folding a closed chain is so hard in general, the problem of protein folding cannot be trivial. Indeed, the solution of the problem has presented great computational difficulties yet to be overcome.

#### 1.4.2. *Biological and physical basis*

Although there may be many factors, including small molecules like chaperones, that affect protein folding, it is agreed that the sequence of amino acids which form the protein determines solely the structure it folds into. In other words, the information that is required for a protein to fold into its structure is fully contained in its constituent amino acids and their order in the sequence. This fact can also be understood to mean that under normal biological or physical conditions, the same sequence of amino acids will fold into the same protein, while other factors may affect only the rate of folding but not the ultimate structure.

The folding of a protein into a 3D structure is a result of the physical interactions among the amino acids in the protein sequence or, more accurately, among the atoms in the protein. The interactions are driven by the potential energy of the protein. The atoms interact to seek a way of lowering the total potential energy. It is assumed that a protein reaches the global potential energy minimum when forming the native structure. The structure is therefore also considered the most stable. Of course, at nonzero temperatures, the free energy instead of potential energy should be used in the assumption, because in general it is the free energy that every physical system tends to minimize.

Depending on the potential energy landscape, folding can be a complicated process. There is a theory that folding takes a unique pathway for each protein, crossing a series of intermediate states. The most recent theoretical and experimental studies have shown that the protein potential energy landscape resembles a funnel shape, with many local minima along the walls of the funnel but the global minimum at the bottom. The local minima along the walls should be easy to skip, while there are still many local minima at the bottom around the global minimum, making the global minimum hard to be found. Also, because of such an energy landscape, there is so far no strong evidence supporting unique folding pathways. Therefore, a common consensus is that folding may have multiple pathways, as it seems possible to start from any point of the funnel and follow a path to reach the bottom of the energy landscape.

### 1.4.3. *Computer simulation*

The process of folding is hard to be observed experimentally. The only properties that can be observed in physical experiments may be the initial and ending structures and the time for folding. Computer simulation is the only way to determine the folding pathway based on physical principles. To make the computation affordable, force field functions have been developed so that the forces for various atomic-level interactions in proteins can be calculated relatively easily and precisely. By using these force functions, we can in principle simulate protein motions such as protein folding by solving a system of equations of motion for any protein of interest.

There are many methods to solve the equations. The basic idea is to start with the initial positions and velocities of the atoms and then try to find the positions and velocities of the atoms at later times. At each time point, the positions and velocities of the atoms are calculated based on their previous positions and velocities, and the time is advanced by a small step. However, the main challenge is that the step size has to be small — in the order of femtoseconds — to achieve the desired accuracy, while protein folding requires milliseconds or even seconds to complete. Clearly, the simulation of folding may take

millions of billions of steps, which can hardly be affordable on even the most powerful computer to date.

Duan and Kollman (1998) performed a 1- $\mu$ s simulation for a small protein, the chicken villin headpiece subdomain HP36. This protein has 36 residues, with 3000 water molecules added into the system. The simulation was carried out on a 256-processor CRAY T3E, using the software AMBER for the calculation of the force field. The CRAY T3E is a very powerful supercomputer. Even with one processor, it can provide much more computing power than a regular workstation. Still, the entire simulation took 100 days to be completed.

The 1- $\mu$ s computer simulation of Duan and Kollman (1998) was able to reveal the major folding pathway of HP36, which usually folds in about 10–100  $\mu$ s. The final structure that the simulation converged to showed a close correlation with the structure obtained through NMR experiments. More importantly, based on the simulation, a folding trajectory was obtained and many intermediate states were discovered. The work was indeed exciting because it was the first protein structure folded on a computer using solely physical principles. It was also the first time that a complete folding pathway was plotted and analyzed.

Efforts have also been made to utilize loosely connected networks of computers for protein folding simulation. Most notable is an Internet website, *folding@home*, developed to perform folding simulation over the Internet. The idea is that over the Internet, there are hundreds of thousands of computers often in an idle state; therefore, *folding@home* organizes available computers to donate time for some simulation tasks. For this purpose, the simulation is conducted as follows.

A set of trajectories is first followed by a group of available computers on the Internet. Once in a while, if a faster folding trajectory is found, a new set of starting points is created around the end point of the trajectory, and the computers are stopped to follow the trajectories started with the new points. The process continues until a folding pathway is found by connecting a sequence of restarted trajectories. During the simulation, a computer can participate at any time and

request an unfinished task. It can return the task to the system whenever the time is up. The returned task may be continued again when another computer becomes available.

The website *folding@home* has run for several years, during which quite a few proteins have been folded on the Internet. For example, the folding of the protein HP36 was simulated by Duan and Kollman (1998) for only 1  $\mu$ s, but the real folding was estimated to take 10–100  $\mu$ s. It turns out that *folding@home* was able to complete the entire simulation and provide a full pathway description for folding the protein. The results from *folding@home* may not yet deliver completely satisfactory answers to the questions of protein folding, but they have shown some promising directions that may lead to an ultimate computational solution to the problem of protein folding if proper models, algorithms, and computing resources can be developed and utilized.

#### 1.4.4. *Alternative approaches*

A computational bottleneck in fold simulation is that a long sequence of iterative steps must be carried out, where each step can start only after the previous one has completely finished. In each step, there are not many calculations to be performed; and even with the most powerful computer (such as a parallel computer), the speedup will be limited and the simulation cannot be completed in a reasonable time. Two alternative approaches to the folding problem are worth mentioning: the boundary value formulation and potential energy minimization approaches.

The first approach, which assumes the availability of the end structure, finds a trajectory that connects the given initial and ending structures. The approach cannot be used if the ending structure is unknown. In such a case where the ending structure is available, the trajectory can be formulated as a boundary value problem, which can be done differently from an initial value problem and in a more parallel fashion than the conventional sequential manner. The latter property makes it possible to perform a complete fold simulation if a massively parallel computer is used.

The second approach sacrifices the attempt to obtain the folding pathway and focuses solely on finding the final folded structure, which is already a very challenging problem. This approach assumes that the native structure of a protein corresponds to the global minimum of the potential energy of the protein. Therefore, the structure may be found directly, using optimization methods instead of dynamics simulation. However, global optimization is difficult, especially for functions with many local minima such as the potential energy functions for proteins. The success of this approach therefore depends on the development of a global optimization algorithm that can be applied effectively to proteins, motivating many intensive investigations along this line.

## Selected Further Readings

### *Protein structure*

- Campbell NA, Reece JB, *Biology*, 7th ed., Benjamin Cummings, 2004.
- Berg JM, Tymoczko JL, Stryer L, *Biochemistry*, W. H. Freeman, 2006.
- Branden CI, Tooze J, *Introduction to Protein Structure*, 2nd ed., Garland Publishing Inc., 1999.
- Lesk AM, *Introduction to Protein Architecture: The Structural Biology of Proteins*, Oxford University Press, 2001.
- Creighton TE, *Proteins: Structures and Molecular Properties*, 2nd ed., Freeman & Co., 1993.
- Jacob-Molina A, Arnold A, HIV reverse transcriptase function relationships, *Biochemistry* 30: 6351–6361, 1991.
- Rodgers DW, Gamblin SJ, Harris BA, Ray S, Culp JS, Hellmig B, Woolf DJ, Debouck C, Harrison SC, The structure of unliganded reverse transcriptase from the human immunodeficiency virus type 1, *Proc Natl Acad Sci USA* 92: 1222–1226, 1995.
- Telesnitsky A, Goff SP, Reverse transcriptase and the generation of retroviral DNA, in *Retroviruses*, Coffin J, Hughes S, Varmus H (eds.), Cold Spring Harbor Laboratory Press, pp. 121–160, 1997.
- Zahn R, Liu A, Lührs T, Riek R, von Schroetter C, Garcia FL, Billeter M, Calzolari L, Wider G, Wüthrich K, NMR solution structure of the human prion protein, *Proc Natl Acad Sci USA* 97: 145–150, 2000.
- Aguzzi A, Montrasio F, Kaeser P, Prions: Health scare and biological challenge, *Nat Rev Mol Cell Biol* 2: 118–125, 2001.

Aguzzi A, Heikenwalder M, Cannibals and garbage piles, *Nature* **423**: 127–129, 2003.

### ***Structure determination***

Woolfson MM, *Introduction to X-ray Crystallography*, 2nd ed., Cambridge University Press, 2003.

Drenth J, *Principles of Protein X-ray Crystallography*, 3rd ed., Springer, 2006.

Wuthrich K, *NMR of Proteins and Nucleic Acids*, Wiley, 1986.

Keeler J, *Understanding NMR Spectroscopy*, Wiley, 2005.

Schlick T, *Molecular Modelling and Simulation: An Interdisciplinary Guide*, Springer, 2003.

Bourne PE, Weissig H, *Structural Bioinformatics*, John Wiley & Sons, Inc., 2003.

### ***Dynamics simulation***

Haile JM, *Molecular Dynamics Simulation: Elementary Methods*, Wiley, 2001.

MaCammon JA, Harvey SA, *Dynamics of Proteins and Nucleic Acids*, Cambridge University Press, 2004.

Brooks III CL, Karplus M, Pettitt BM, *Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics*, Advances in Chemical Physics, Vol. 71, Wiley, 2004.

Cui Q, Bahar I, *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*, Chapman & Hall/CRC, 2005.

### ***Protein folding***

Dobson CM, Fersht AR, *Protein Folding*, Cambridge University Press, 1996.

Pain RH, *Mechanism of Protein Folding*, Oxford University Press, 2000.

Wolynes PG, Folding funnels and energy landscapes of larger proteins within the capillarity approximation, *Proc Natl Acad Sci USA* **94**: 6170–6175, 1997.

Mirny L, Shkhovich E, Protein folding theory: From lattice to all-atom models, *Annu Rev Biophys Biomol Struct* **30**: 361–396, 2001.

- Duan Y, Kollman P, Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution, *Science* **282**: 740–744, 1998.
- Duan Y, Kollman PA, Computational protein folding: From lattice to all-atom, *IBM Syst J* **40**: 297–309, 2001.
- Shirts MR, Pande VS, Screen savers of the world unite!, *Science* **290**: 1903–1904, 2000.
- Pande VS, Baker I, Chapman J, Elmer SP, Khaliq S, Larson SM, Rhee YM, Shirts MR, Snow CD, Sorin EJ, Zagrovic B, Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing, *Biopolymers* **68**: 91–109, 2002.
- Elber R, Meller J, Olender R, Stochastic path approach to compute atomically detailed trajectories: Application to the folding of C peptide, *J Phys Chem* **103**: 899–911, 1999.
- Wales DJ, Scheraga HA, Global optimization of clusters, crystals, and biomolecules, *Science* **285**: 1368–1372, 1999.