Chapter 10
Foundations for the Study of Structure and Function of Proteins

Zhirong Sun

10.1 Introduction

Proteins are the most abundant biological macromolecules, occurring in all cells and all parts of cells. Moreover, proteins exhibit enormous diversity of biological function and are the most final products of the information pathways. Protein is a major component of protoplasm, which is the basis of life. It is translated from RNA and composed of amino acid connected by peptide bonds. It participates in a series of complicated chemical reactions and finally leads to the phenomena of life. So we can say it is the workhorse molecule and a major player of life activity. Biologists focus on the diction of structure and function of proteins by the study of the primary, secondary, tertiary, and quaternary dimensional structures of proteins, posttranscriptional modifications, protein-protein interactions, the DNA-proteins interactions, and so on.

10.1.1 Importance of Protein

DNA, RNA, proteins, etc. are the basic components of life. DNA is the vector of genetic information and is transcribed into RNA which is in turn translated into protein. Protein is the expression of genetic information, the performer of kinds of biological functions, and the sustainer of metabolic activities in the organisms. Protein plays an important role in the whole processes of life, including the appearance of life to the growth of life to apoptosis.

There are two examples illustrating the importance of protein. The first one is about the SARS. One protein is found to increase the self-copy efficiency for 100

Z. Sun (✉)
School of Life Sciences, Tsinghua University, Beijing 100084, China
e-mail: sunzhr@mail.tsinghua.edu.cn

R. Jiang et al. (eds.), *Basics of Bioinformatics: Lecture Notes of the Graduate Summer School on Bioinformatics of China*, DOI 10.1007/978-3-642-38951-1_10,

times or so, which makes the viruses propagate at a high rate. Another example is a flu virus protein whose structure looks like a narrow-neck bag. This strange structure of the protein can help the virus resist drugs.

10.1.2 Amino Acids, Peptides, and Proteins

All proteins, whether from the most ancient lines of bacteria or from the most complex forms of life, are constructed from the same ubiquitous set of 20 amino acids, covalently linked in characteristic linear sequences. Proteins are the polymers of 20 amino acids. Different combinations of these 20 amino acids result in varied structures and functions of proteins. Protein structures are studied at primary, secondary, tertiary, and quaternary levels. Proteins have widely diverse forms and functions, including enzymes, hormones, antibodies, transporters, muscle, lens protein of eyes, spider webs, rhinoceros horn, antibiotics, and mushroom poisons.

10.1.2.1 Protein Research: What to Study and How to Study?

What should we study? What is the core problem? How can we study? What is most remarkable is that cells can produce proteins with strikingly different properties and activities by joining the same 20 amino acids in many different combinations and sequences. Nowadays, biologists study protein from these aspects: structure and function, the transfer of information.

10.1.2.2 Amino Acid

All 20 standard amino acids found in proteins are α-amino acids. Figure 10.1 shows the structure formula of α-amino acids. Each amino acid has a different side chain (or R group, R = "remainder of the molecule") and is given a three-letter abbreviation and a one-letter symbol. Biologists often use the first three letters or the first letter. The 20 amino acids of proteins are often referred to as the standard amino acids. All proteins in all species (from bacteria to human) are constructed from the same set of 20 amino acids. All proteins, no matter how different they are in structure and function, are made of the 20 standard amino acids. Figure 10.1 shows the structure formulae of all the 20 amino acids.

10.1.2.3 Protein Structure Hierarchy

Protein structures are studied at primary, secondary, tertiary, and quaternary levels. There are tight correlations among these levels.

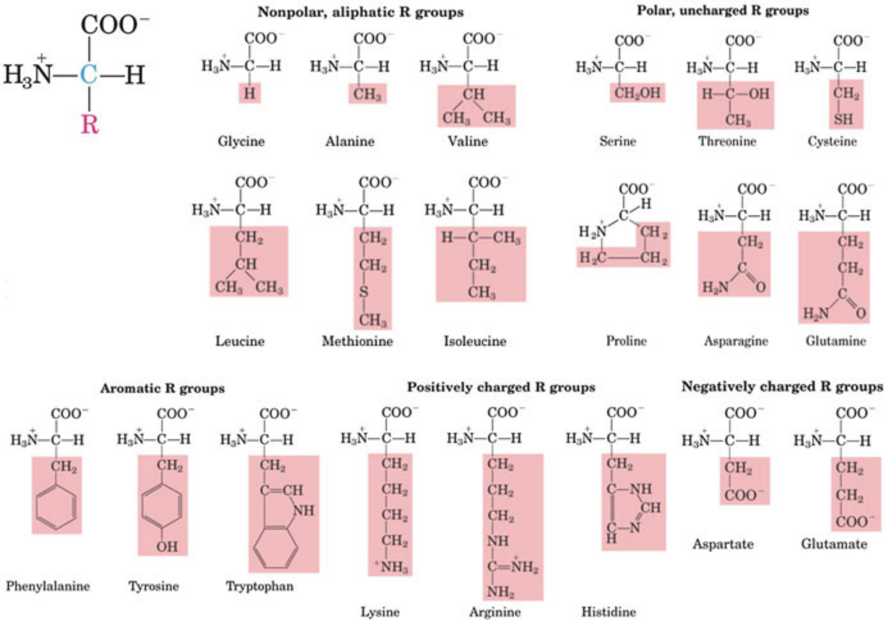


Fig. 10.1 Structure formulae of all the 20 amino acids

10.1.2.4 Different Classes of Proteins

From the aspect of chemical structures of proteins, proteins can be classified into two classes. If proteins are completely composed of amino acids, these proteins are called simple proteins, such as insulin; if there are other components, they are named conjugated proteins like hemoglobin.

According to the symmetry of proteins, proteins can be divided into globin and fibrin. Globins are more symmetric and similar to balls or ovals in shape. Globins dissolve easily and can crystallize. Most proteins are globins. Comparatively, fibrins are less symmetric and look like thin sticks or fibers. They can be divided into soluble fibrins and unsolvable fibrins.

Simple proteins can be subdivided into seven subclasses: albumin, globulin, glutelin, prolamine, histone, protamine, and scleroprotein. Conjugated proteins can also be subdivided into nucleoprotein, lipoprotein, glycoprotein and mucoprotein, phosphoprotein, hemoprotein, flavoprotein, and metalloprotein. Different classes of proteins have various functions. These include serving as:

1. Catalyzers of metabolism: enzyme
2. Structural component of organisms
3. Storage component of amino acid
4. Transporters
5. Movement proteins
6. Hormonal proteins

7. Immunological proteins
8. Acceptor and for transfer of information
9. Regulatory or control mechanisms for the growth, division, and the expression of genetic information

10.1.3 *Some Noticeable Problems*

Biological function and biological character are two different concepts. Characters can be shown from a chemical reaction, while functions of molecules are shown by the whole system in several cooperated reactions. Functions are related to the molecule interactions.

10.2 Basic Concept of Protein Structure

10.2.1 *Different Levels of Protein Structures*

10.2.1.1 The Basic Unit of Protein (Fig. 10.2)

10.2.1.2 Polypeptide Chain

Peptide and Peptide Bond

A peptide bond is made up by connecting an α-COOH of an amino acid and the α-NH_3 (Figs. 10.3 and 10.4). The simplest peptide composed of two amino acids is called dipeptide, containing one peptide bond. Those containing three, four, and five peptide bonds are called tripeptide, tetrapeptide, and pentapeptide, respectively. The peptide chain loses a molecule of H_2O when forming a peptide bond. In a polypeptide, an amino acid unit is called a residue.

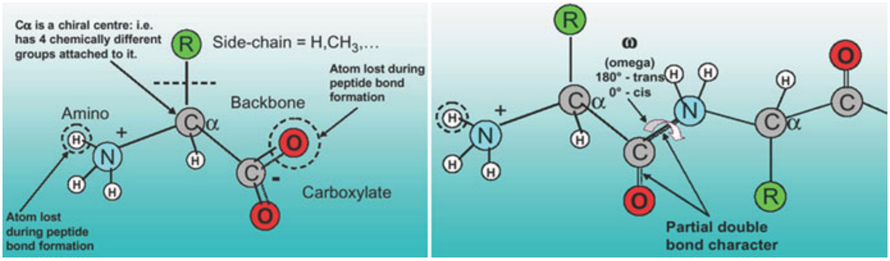

Cα is a chiral centre: i.e. has 4 chemically different groups attached to it.
R
Side-chain = H,CH3,...
Atom lost during peptide bond formation
Backbone
Amino
Cα
+
N
H
C
O
-
Carboxylate
Atom lost during peptide bond formation
R
ω
(omega)
180° - trans
0° - cis
N
Cα
C
O
R
Partial double bond character

Fig. 10.2 Common structure of amino acid (*left*) and formation of polypeptide chain (*right*)

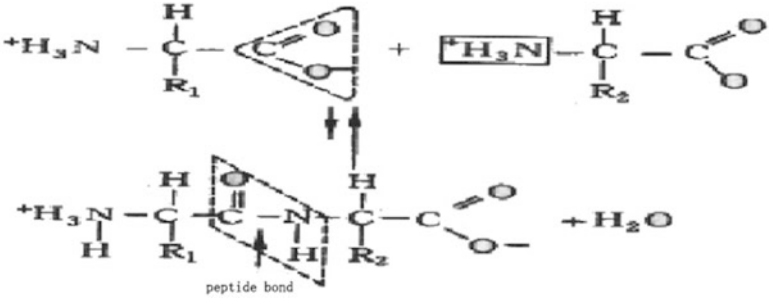


Fig. 10.3 The formation of a peptide bond

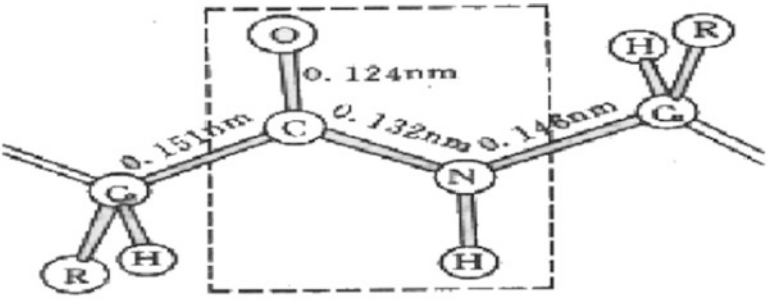


Fig. 10.4 The rigidity and coplanarity of peptide bond

The Fold of Polypeptide Chain and Dihedral Angle

The repeated structure on the backbone of polypeptide is called peptide unit or planar unit of peptide. Peptide bond cannot turn freely because of its double-bond character. The bonds beside peptide unit can wheel freely, which are described using dihedral angles ϕ and ψ.

10.2.1.3 The Imagination of a Polypeptide Chain (Fig. 10.5)

10.2.1.4 The Peptide Chain Is Directional

1. An amino acid unit in a peptide chain is called a residue.
2. The end having a free α-amino group is called amino-terminal or N-terminal.
3. The end having a free α-carboxyl group is called carboxyl-terminal or C-terminal.
4. By convention, the N-terminal is taken as the beginning of the peptide chain and put at the left (C-terminal at the right). Biosynthesis starts from the N-terminal.

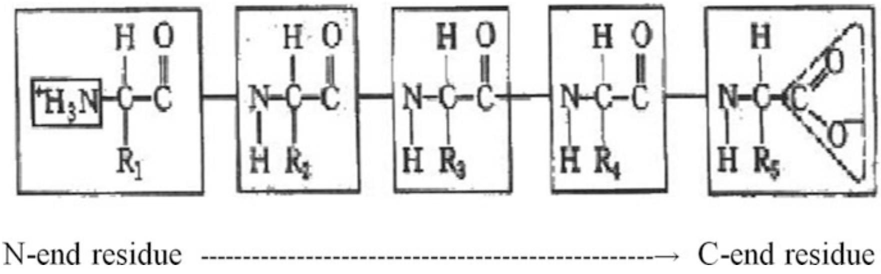


Fig. 10.5 The structure of a pentapeptide

10.2.2 *Acting Force to Sustain and Stabilize the High-Dimensional Structure of Protein*

10.2.2.1 The Interaction of Biological Molecules

The Electronic Interaction of Biological Molecules

The electronic interaction includes charge-charge interaction, charge-dipole interaction, dipole-dipole interaction, and induced dipole interaction.

Dipole moment

$$\mu = g \cdot l,\ u = -\mu \cdot E$$

Charge-charge interaction

Charge-dipole interaction (Fig. 10.6)

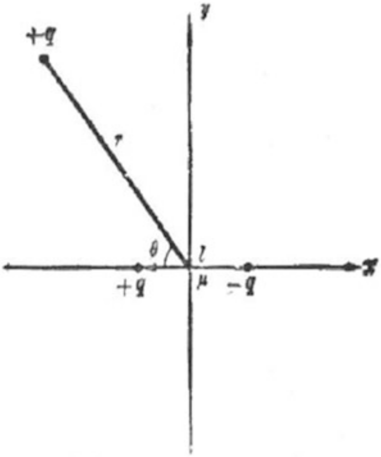
Dipole-Dipole Interaction

When the radius vector between two dipoles and the center is far bigger than the length of dipoles, namely, $r \gg 1$, the interaction of these two dipoles is:

$$U = \frac{1}{\varepsilon r^3}\left[\mu_{\mathrm{A}} \cdot \mu_{\mathrm{B}} - \frac{3\left(\mu_{\mathrm{A}} \cdot r\right)\left(\mu_{\mathrm{B}} \cdot r\right)}{r^2}\right] \quad l \ll r$$

Induced dipoles

Fig. 10.6 The interaction of a positive charge and dipole μ



The neutral molecules or groups with overlapped positive and negative charges will be polarized by electric field and become induced dipoles. The dipole moment:

$$\mu_{\text{ind}} = -a_{\text{ind}}E$$

The Hydration of Polar Groups

Hydration is the process of the subject interacting or combining with water.

10.2.2.2 The Force to Sustain and Stabilize the High-Dimensional Structure of Proteins

The forces that sustain the structure of proteins are the so-called weak interaction, non-covalent bond, or inferiority bond, including hydrogen bond, hydrophobic interaction, electrostatic interaction, and van der Waals force. When these weak interactions present independently, they are weak bond, but when these bonds are added together, a strong force will form to sustain the protein structure space.

Electrostatic Force

Under the physiological condition, the side chain of acidic amino acid can be broken down into negative ions, while the side chain of basic amino acid can be disassociated into positive ions. Some atoms will form dipoles because of polarization. These interaction forces between charges or dipoles are called electrostatic force and it meets the Coulomb's law.

Van der Waals Force

Van der Waals force can also be called van der Waals bond. It includes attractive force and repulsion force. Van der Waals attractive force is in inverse ratio to the sixth power of the distance between atoms or groups. When they are too close to each other, they will repel each other. The van der Waals bond length is 0.3–0.5 nm. The bond energy is 1–3 kcal/mol.

Although the van der Waals force is weak, when the surfaces of two big molecules are close enough to each other, this force is very important. It contributes to sustain the tertiary structure and quaternary structure.

10.3 Fundamental of Macromolecules Structures and Functions

10.3.1 Different Levels of Protein Structure

Protein structures have conventionally been understood at four different levels (Fig. 10.7):

1. The primary structure is the amino acid sequence (including the locations of disulfide bonds).
2. The secondary structure refers to the regular, recurring arrangements of adjacent residues resulting mainly from hydrogen bonding between backbone groups, with α-helices and β-pleated sheets as the two most common ones.
3. The tertiary structure refers to the spatial relationship among all amino acid residues in a polypeptide chain, that is, the complete three-dimensional structure.
4. The quaternary structure refers to the spatial arrangements of each subunit in a multi-subunit protein, including nature of their contact.

10.3.1.1 The Formation of Protein Structure Level and the Folding of Peptide Chain

In protein solution, if the environment changes, for example, pH, ion strength, or temperature changes, the natural structure of protein may disintegrate and leads to the denaturation of proteins. This process is called protein denaturation. When the condition is normal, if the denatured protein can have their natural structure and character back, then the protein will renature.

The way to make bean curd by heating the solution of bean protein and adding a little salt is an example to make use of the protein denaturation to deposit protein.

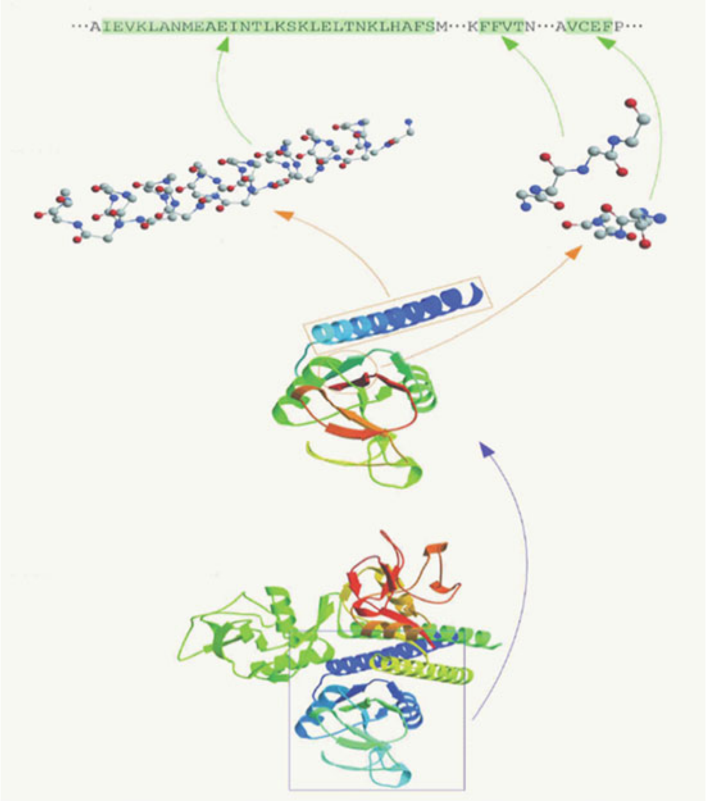


Fig. 10.7 Different levels of protein structure

Table 10.1 The nucleic and protein databases

Nucleic database	Protein database
EMBL	SWISS-PROT
GenBank	PIR
DDBJ	MIPS
	TrEMBL
	NRL-3D

10.3.2 Primary Structure

According to the classical view, the primary structure of protein decides the high-level structure of proteins. So the high-level structure can be inferred from the primary structure. We can align multiple protein sequences (Table 10.1, Figs. 10.8 and 10.9).

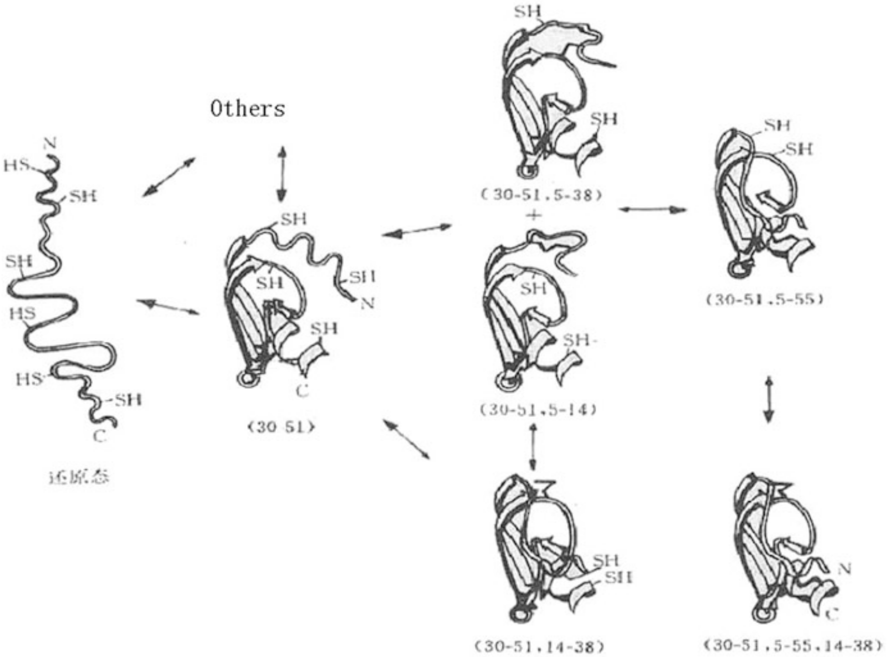


Fig. 10.8 The process of BPT1 folding from loose peptide chains into correct active tertiary structure

In the 1980s, when sequences started to accumulate, several labs saw advantages to establishing central repositories. The trouble is many labs thought the same and made their own. The proliferation of databases causes problems. For example, do they have the same format? Which one is the most accurate, up-to-date, and comprehensive? Which one should we use?

10.3.3 Secondary Structure

10.3.3.1 Various Kinds of Protein Secondary Structure

Local organization of protein backbone is α-helix, β-strand (which assembles into β-sheet), turn, and interconnecting loop.

α-Helix is in a shape of stick. Tightly curled polypeptide backbone forms the inner side of the stick; the side chains expand outside in the form of helix. α-Helix tends to be stable because the hydrogen in NH and the oxygen in the fourth residue CO form hydrogen bond. Each helix contains 3.6 residues. The helix distance is 0.54 nm.

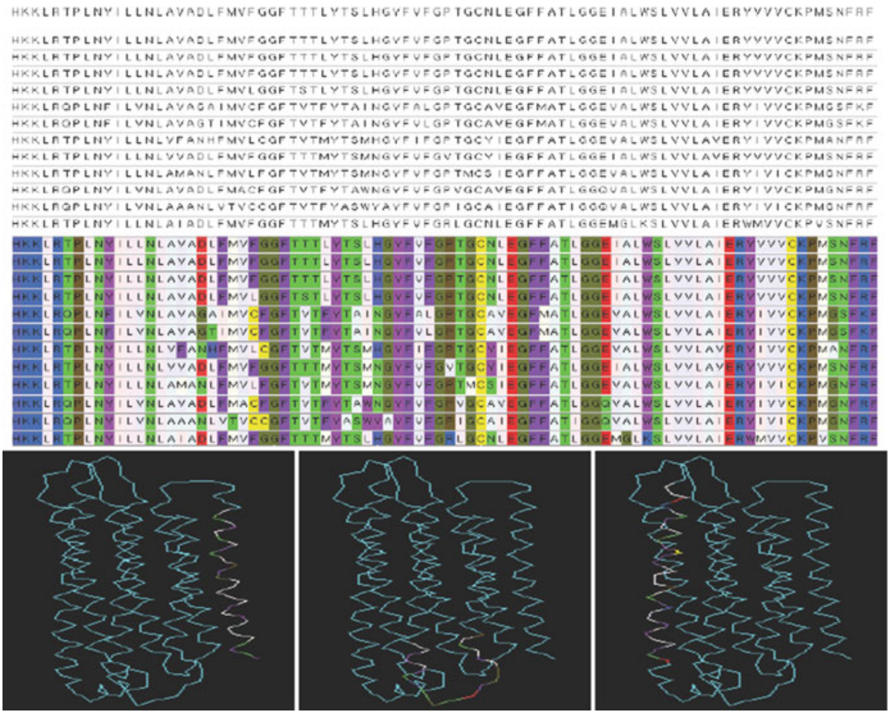

HKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLGGEIALWSLVVLAIERYVVVCKPMSNFRF

Fig. 10.9 An alignment of protein primary structure

β-Sheet is another frequently occurrence structure. Two or more fully expended polypeptides cluster together laterally. Hydrogen bond is formed by –NH and C=O on the neighboring peptide backbones. These polypeptide structures are β-sheet. In the β-sheets, all peptides join in the cross-linking between hydrogen bonds. The hydrogen bonds are almost vertical to the long axis of peptide chains. Along the long axis of peptide chain, there are repeated units.

β-Sheet includes two types. One is the parallel sheet. The arrangement polarization of its peptide chain (N–C) is unidirectional. The N-end of all the peptide chains is in the same direction. Another one is antiparallel. The polarization of the peptide chain is opposite for the neighboring chains.

In the backbones of polypeptide chain, the structures which are different from the α-helix and β-sheet are called random coil. Random coils mean the irregular peptide chain. For most globins, they often contain a great amount of random coils besides α-helix and β-sheet. In random coils, β-turn is a very important structure.

β-Turn can also be called reverse turn, β-bend, and hairpin structure. It is composed of four successive amino acids. In this structure, the backbone folds in a degree of 180°. The oxygen on C=O of the first residue and hydrogen on the N–H

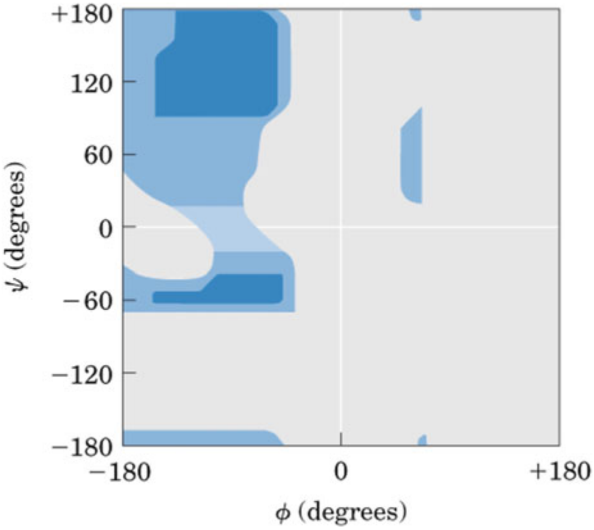


Fig. 10.10 Ramachandran plot for L-Ala residues. *Dark blue* area reflects conformations that involve no steric overlap and thus are fully allowed; *medium blue* indicates conformations allowed at the extreme limits for unfavorable atomic contacts; the *lightest blue* area reflects conformations that are permissible if a little flexibility is allowed in the bond angles (Color figure online)

of the fourth residue form hydrogen bond. The structure of the β-turn is determined by the dihedral angel (ϕ_2, ψ_2; ϕ_3, ψ_3) made of the second residue and third residue.

10.3.3.2 The Variability of Protein Secondary Structure (Figs. 10.10, 10.11, and 10.12)

10.3.4 Supersecondary Structure

Two or several secondary structure units connected by connecting peptides can form special space structures. They are called protein supersecondary structures.

10.3.4.1 High-Frequency Supersecondary Structure Motif

Protein databases (PDB)

1. Analysis of main-chain conformations in known protein structure.
2. 12,318 residues from 84 proteins, structure 5,712 fell outside the regions of regular structure.
3. Torsion angles (φ,ψ) for the 5,712 residues were calculated and allocated to seven classes in the Ramachandran plot: a,b,e,g,l,p,t ⇒ a,b,e,l,t H, E.

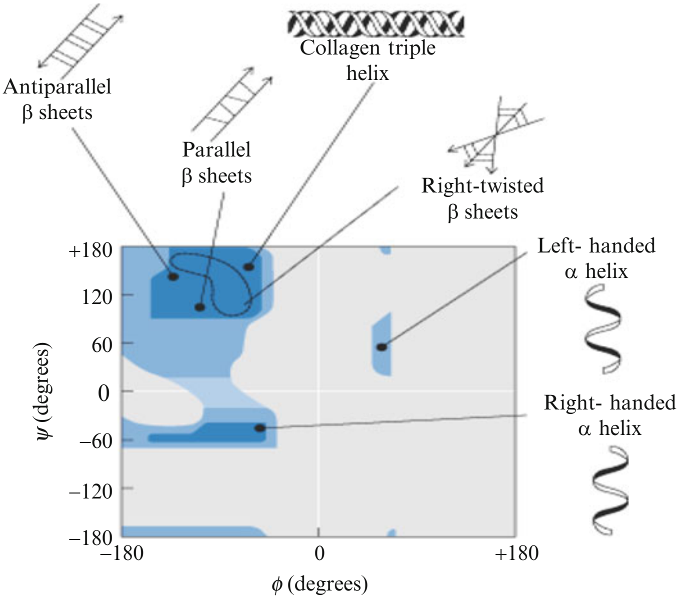


Fig. 10.11 Ramachandran plots for a variety of structures

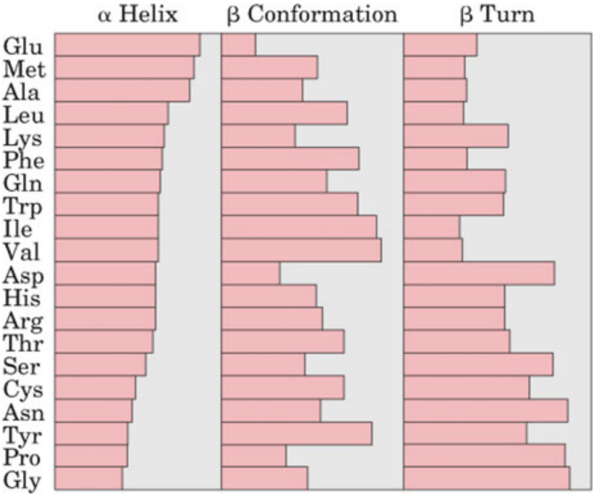


Fig. 10.12 Relative probabilities that a given amino acid will occur in the three common types of secondary structure

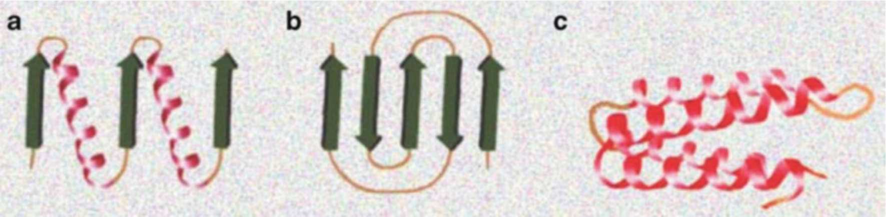


Fig. 10.13 (**a**) Rossmann fold; (**b**) Greek key topology structure; (**c**) four-helix bundle

4. Statistical analysis of lengths and conformations of the supersecondary motif. Sequence and conformational data were stored for three successive residues in each of the two elements of secondary structure on either side of the connecting peptide, for example, HHH abl EEE.
5. Classification of the pattern and conformation for supersecondary structure motifs.

10.3.4.2 Basic Protein Supersecondary Structure Motif

α-α-Hairpin is made up of two almost antiparallel α-helixes connected by a short peptide. This short peptide is usually composed of 1–5 amino acids.

β-β-Hairpin is made up of two antiparallel β-sheets connected by a short peptide. This peptide is usually composed of 1–5 amino acids.

α-α-Corner is made up of α-helixes on two different planes connected by a connecting peptide. The vector angle between these two α-helixes is nearly right angle.

α-β-Arch structure is made up of an α-helix and a β-sheet connected by a short peptide. The most frequently occurring α-β-structure is composed of three parallel β-sheets and two α-helixes. This structure is called Rossmann sheet.

10.3.4.3 Characteristic Description of Protein Supersecondary Structure

There are mainly three characteristic descriptions of supersecondary structures: sequence pattern, hydrophobic pattern, and H-bond pattern.

10.3.4.4 Complicated Supersecondary Structure Motif

In protein structures, many basic supersecondary structure motifs form some more complicated complexes motif, which are called complicated supersecondary structures.

The commonly occurring complicated supersecondary structures include Rossmann fold (Fig. 10.13a), Greek Key topology structure (Fig. 10.13b), and four-helix bundle (Fig. 10.13c), etc. (Figs. 10.14 and 10.15)

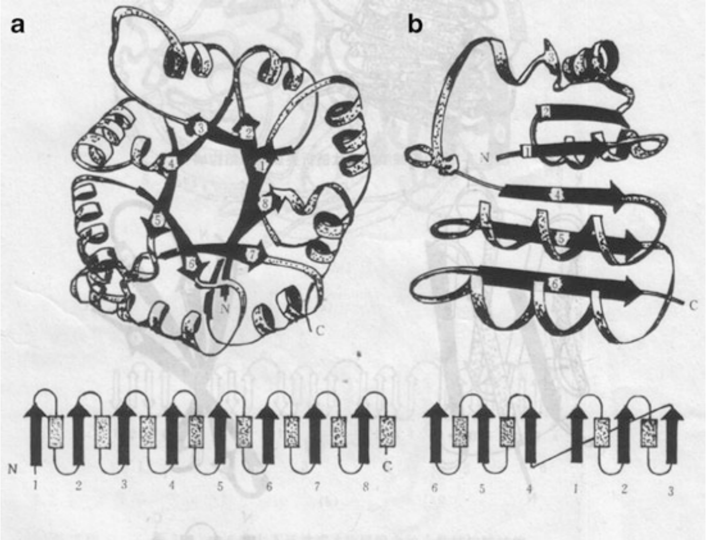


Fig. 10.14 Two different α/β-domains frequently observed in many proteins (**a**) closed β-barrel and (**b**) open curled β-sheet

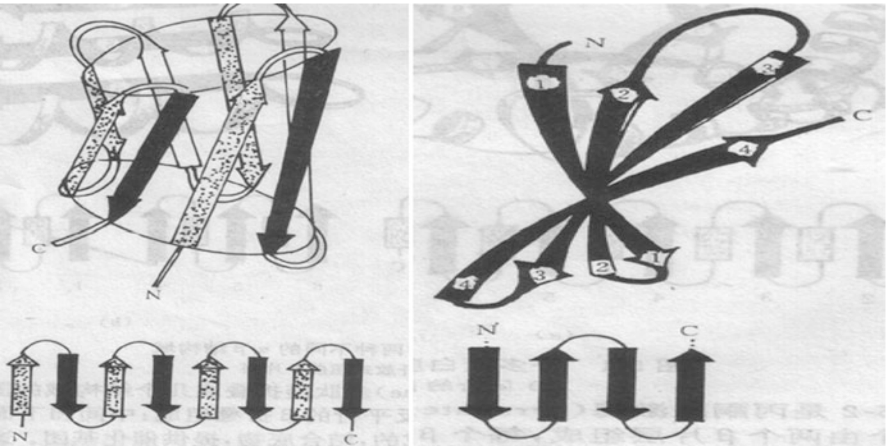


Fig. 10.15 Up-down β-barrel structure and up-down open β-sheet structure

10.3.4.5 Protein Tertiary Structure

Polypeptide chains further fold by non-covalent bond interaction and curl into more complicated configuration, which is called tertiary structure.

For bigger protein molecules, polypeptide chains are always composed of two or more independent three-dimensional entity. These entities are called domains.

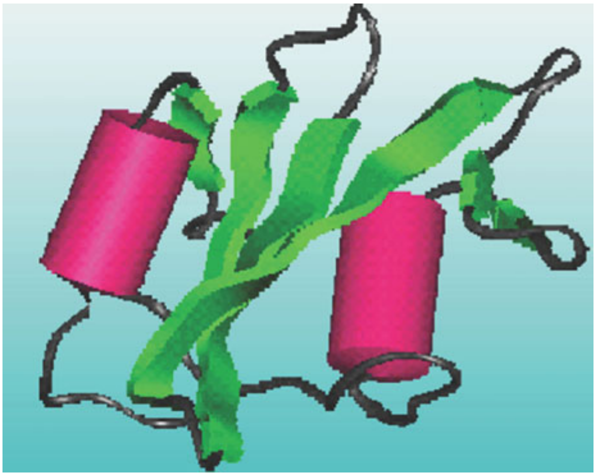


Fig. 10.16 Tertiary structure

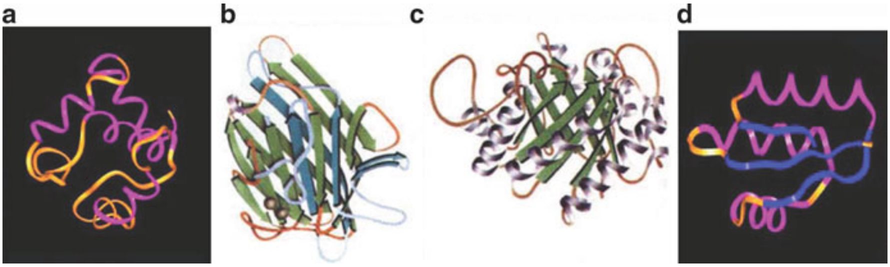


Fig. 10.17 (**a**) α-Protein, (**b**) β-protein, (**c**) α + β-protein, and (**d**) α/β-protein

According to the amount of α-helix and β-sheet, proteins can be divided into four types: α-protein, β-protein, α + β-protein, and α/β-protein (Fig. 10.16).

α-Protein contains more than 40 % of α-helix and less than 10 % of β-sheet (Fig. 10.17a). β-Protein contains more than 40 % of β-sheet and less than 10 % of α-protein (Fig. 10.17b). α + β-Protein contains more than 10 % of α-helix and β-sheet. α,β-Clusters in different regions. α/β-Protein (Fig. 10.17c) contains more than 10 % of α-helix and β-sheet. These two configurations appear in the peptide chain alternatively. The two configurations of different α/β-proteins (Fig. 10.17d) arrange face to face. The shape of the whole molecule varies a lot.

10.3.4.6 Protein Quaternary Structure

Spatial arrangement of subunits in a protein that contains two or more polypeptide chains is called quaternary structure. It often involves symmetry, but doesn't have to.

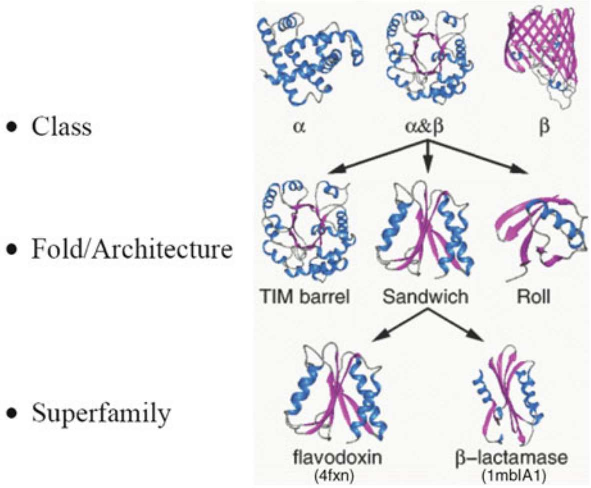


Fig. 10.18 The hierarchy of structural classification

Subunits of proteins form quaternary structure by hydrophobic interaction, H-bond, and van der Waals. The number of most oligomeric proteins is even. There are always one or two types of subunits. The arrangement of most oligomeric protein molecules is symmetric.

Some globins contain two or more polypeptide chain. These polypeptide chains interact with each other, and each of them has their own tertiary structure. These polypeptide chains are subunits of proteins. From the view of structure, subunit is the smallest covalent unit of proteins. Proteins clustered by subunits are called oligomeric proteins. Subunit is the function unit of oligomeric proteins.

10.3.5 Folds

10.3.5.1 Structural Classification of Protein Structure

The hierarchy of structural classification (Fig. 10.18):

- Class
  - Similar secondary structure content
  - All α, all β, α + β, α/β, etc.
- Folds (architecture)
  - Core structure similarity
  - SSEs in similar arrangement

- Superfamily (topology)
  - Probable common ancestry
- Family
  - Clear evolutionary relationship
  - Sequence similarity >25 %
- Individual protein

There are some databanks of structural classification:

- SCOP
  - Murzin AG, Brenner SE, Hubbard T, and Chothia C.
  - Structural classification of protein structures.
  - Manual assembly by inspection.
  - All nodes are annotated (e.g., all α, α/β).
  - Structural similarity search using 3dSearch (Singh and Brutlag).
- CATH
  - Dr. C.A. Orengo, Dr. A.D. Michie, etc.
  - Class-architecture-topology-homologous superfamily.
  - Manual classification at architecture level.
  - Automated topology classification using the SSAP algorithms.
  - No structural similarity search.
- FSSP
  - L.L. Holm and C. Sander.
  - Fully automated using the DALI algorithms (Holm and Sander).
  - No internal node annotations.
  - Structural similarity search using DALI.
- Pclass
  - A. Singh, X. Liu, J. Chang, and D. Brutlag.
  - Fully automated using the LOCK and 3dSearch algorithms.
  - All internal nodes automatically annotated with common terms.
  - JAVA-based classification browser.
  - Structural similarity search using 3dSearch.

10.3.5.2 Hierarchy of Structure

Homologous family: evolutionarily related with a significant sequence identity
Superfamily: different families whose structural and functional features suggest common evolutionary origin

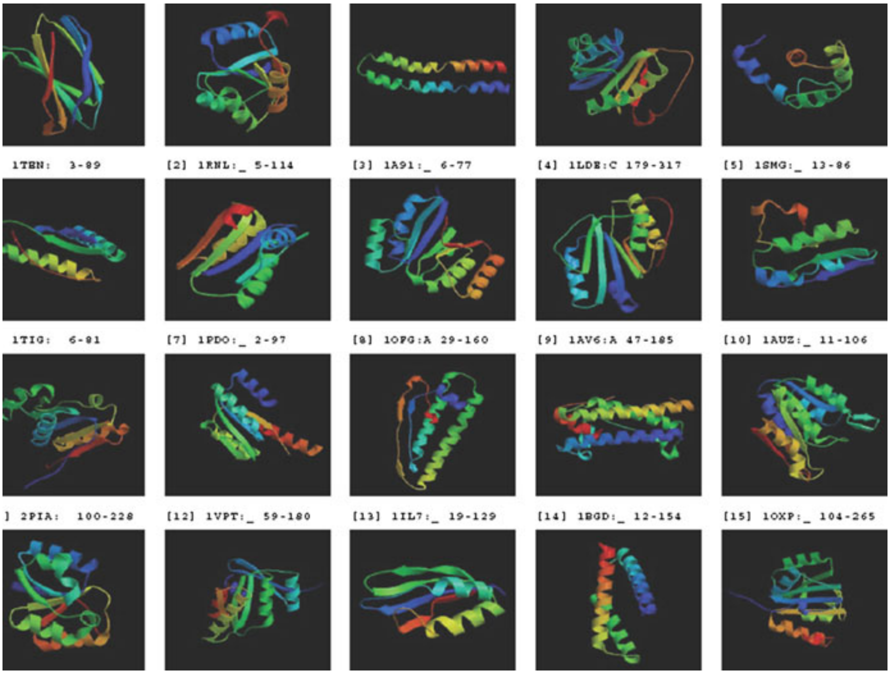


Fig. 10.19 Twenty most frequent common domains (folds)

Folds: different superfamilies having the same major secondary structures in the same arrangement and with the same topological connections (energetic favoring certain packing arrangements)
Class: secondary structure composition

10.3.5.3 Protein Molecule Movement and Function

Proteins have varieties of movements. Movement and structures are the basic elements of protein functions. Protein movement includes short-time and small-amplitude movement, median-time and median-amplitude movement, and long-time and big-amplitude movement (Fig. 10.19).

10.3.6 Summary

Five schemes of protein three-dimensional structures:

1. The three-dimensional structure of a protein is determined by its amino acid sequence.
2. The function of protein depends on its structure.

3. An isolated protein has a unique or nearly unique structure.
4. The most important forces stabilizing the specific structure of a protein are noncovalent interactions.
5. Amid the huge number of unique protein structures, we can recognize some common structural patterns to improve our understanding of protein architecture.

10.4 Basis of Protein Structure and Function Prediction

10.4.1 Overview

In the following part, we are going to talk about the comparative modeling, inverse folding, ab initio, secondary structure prediction, supersecondary structure prediction, structure-type prediction, and tertiary structure prediction.

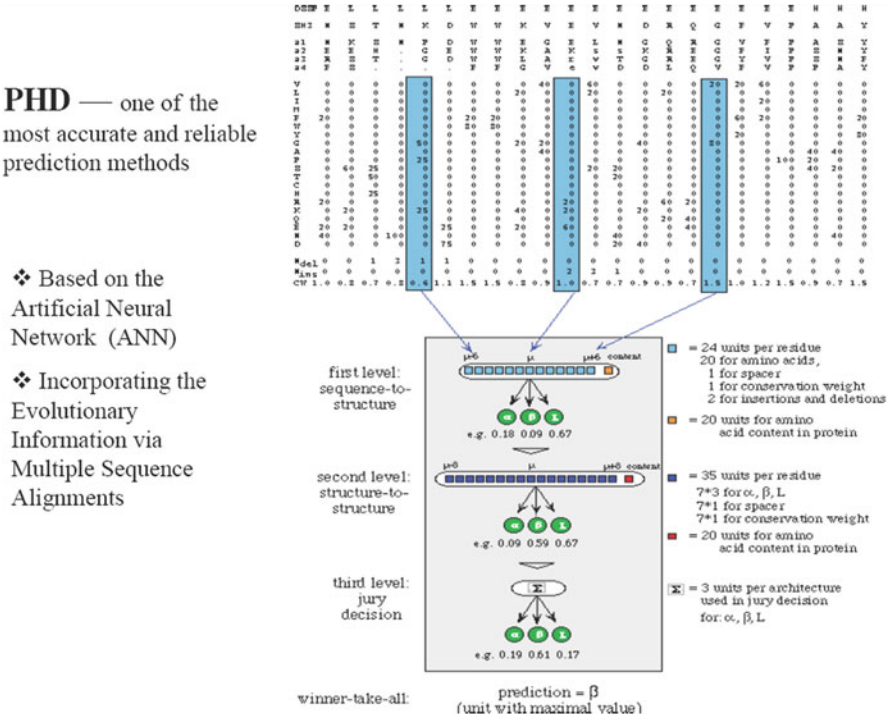
10.4.2 The Significance of Protein Structure Prediction

The development and research of life science show that protein peptide chain-folding mechanism is the most important problem to be solved. How does protein fold from primary structure into active natural tertiary structure is waiting to be answered. The elucidation of the protein peptide chain-folding mechanisms is called decoding the second biological code.

As the human genome and other species genome sequencing plan start and finish, the capacity of databases (e.g., SWISS-PROT) collecting protein sequence increases exponentially. Meanwhile, the capacity of databases (e.g., PDB) collecting protein tertiary crystal structures increases slowly. The increasing rate of the protein sequence number is much greater than that of the known protein structure number. So we need the computational predictive tools to narrow the widening gap.

In the most genome era, one of the biggest challenges we face is to discover the structure and function of every protein in the genome plan. So, predicting protein structure theoretically becomes one way to decrease the disparity between protein structure and sequence.

Why should we predict secondary structure? Because it is an easier problem than 3D structure prediction (more than 40 years of history) and accurate secondary structure prediction can be important information for the tertiary structure prediction. Ever since the first work of prediction of secondary structure done by Chou-Fasman, it has been 30 years. The accuracy is around 60 %. Since 1990s, several machine learning algorithms have been successfully applied to the prediction of protein secondary structure and the accuracy reaches 70 %. From this, we can see a good method can help improve the prediction result significantly.



PHD — one of the most accurate and reliable prediction methods

❖ Based on the Artificial Neural Network (ANN)

❖ Incorporating the Evolutionary Information via Multiple Sequence Alignments

first level: sequence-to-structure
e.g. 0.18 0.09 0.67
second level: structure-to-structure
e.g. 0.09 0.59 0.67
third level: jury decision
e.g. 0.19 0.61 0.17
winner-take-all:
prediction = β (unit with maximal value)
= 24 units per residue 20 for amino acids, 1 for spacer 1 for conservation weight 2 for insertions and deletions
= 20 units for amino acid content in protein
= 35 units per residue 7*3 for α, β, L 7*1 for spacer 7*1 for conservation weight
= 20 units for amino acid content in protein
Σ = 3 units per architecture used in jury decision for: α, β, L
μ-6
μ
μ+6
content

Fig. 10.20 PHD method

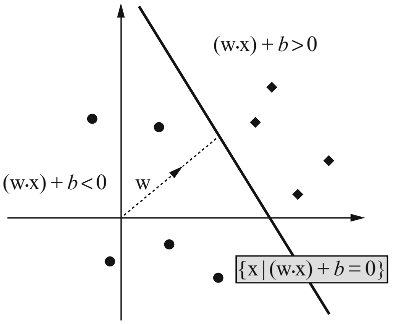
There are a few prediction methods including statistical method (Chou-Fasman method, GOR I-IV), nearest neighbors (NNSSP, SSPAL, Fuzzy-logic-based method), neural network (PHD (Fig. 10.20), Psi-Pred, J-Pred), support vector machine (SVM), and HMM.

10.4.3 The Field of Machine Learning

10.4.3.1 Support Vector Machine

There are many researches in this field. V. Vapnik [1] developed a promising learning theory (Statistical Learning Theory (SLT)) based on the analysis of the nature of machine. Support vector machine (SVM) is an efficient implementation of SLT. SVM has been successfully applied to a wide range of pattern recognition problems, including isolated handwritten digit recognition, object recognition, speaker identification, and text categorization.

Fig. 10.21 The linearly separable case



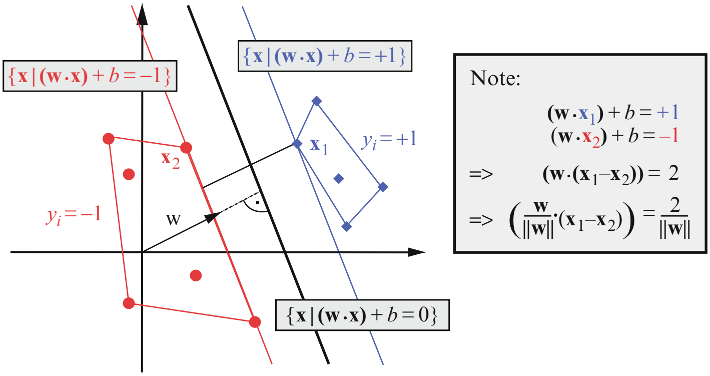

$\{\mathbf{x} \mid (\mathbf{w}{\cdot}\mathbf{x}) + b = -1\}$
$\{\mathbf{x} \mid (\mathbf{w}{\cdot}\mathbf{x}) + b = +1\}$
$\mathbf{x}_2$
$\mathbf{x}_1$
$y_i = +1$
$y_i = -1$
w
$\{\mathbf{x} \mid (\mathbf{w}{\cdot}\mathbf{x}) + b = 0\}$
Note:
$(\mathbf{w}{\cdot}\mathbf{x}_1) + b = +1$
$(\mathbf{w}{\cdot}\mathbf{x}_2) + b = -1$
=> $(\mathbf{w}{\cdot}(\mathbf{x}_1 - \mathbf{x}_2)) = 2$
=> $\left(\frac{\mathbf{w}}{\|\mathbf{w}\|}{\cdot}(\mathbf{x}_1 - \mathbf{x}_2)\right) = \frac{2}{\|\mathbf{w}\|}$

Fig. 10.22 Optimal separating hyperplane (OSH)

For the linearly separable case (Fig. 10.21), the SVM tries to look for one unique separating hyperplane, which is maximal in the margin between the vectors of the two classes. This hyperplane is called Optimal Separating Hyperplane (OSH) (Fig. 10.22).

Introducing Lagrange multipliers and using the Karush-Kuhn-Tucker (KKT) conditions and the Wolfe dual theorem of optimization theory, the SVM training procedure amounts to solving a convex quadratic programming problem:

$$\text{Maximize} \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \cdot y_i y_j \cdot \vec{\mathbf{x}}_i \cdot \vec{\mathbf{x}}_j$$

$$\text{subject to} \quad \alpha_i \geq 0$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \quad i = 1, 2, \ldots, N$$

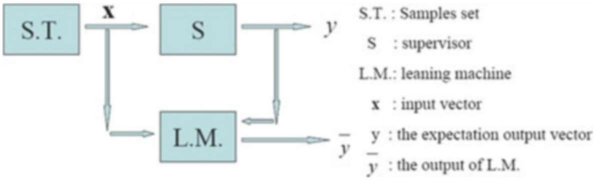


Fig. 10.23 Setting of the machine learning problem. A model of learning from samples. During the learning process, the learning machine observes the pairs (**x**,**y**) (the training set). After training, the machine must on any given *x* return a value. The goal is to return a value which is close to the supervisor's response *y*

The solution is a unique globally optimized result which can be shown to have an expansion (Fig. 10.23):

$$\vec{\mathbf{w}} = \sum_{i=1}^{N} y_i \alpha_i \cdot \vec{\mathbf{x}}_i$$

When an SVM is trained, the decision function can be written as:

$$f\left(\vec{\mathbf{x}}\right) = \operatorname{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i \cdot \vec{\mathbf{x}} \cdot \vec{\mathbf{x}}_i + b\right)$$

For the linearly non-separable case, the SVM performs a nonlinear mapping of the input vectors from the input space R^d into a high-dimensional feature space H and the mapping is determined by a kernel function. Then like the linearly separable case, it finds the OSH in the higher-dimensional feature space H.

The convex quadratic programming problem:

$$\text{Maximize} \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \cdot y_i y_j \cdot K\left(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j\right)$$

$$\text{subject to} \quad 0 \le \alpha_i \le C$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \quad i = 1, 2, \ldots, N$$

The decision function:

$$f(\vec{\mathbf{x}}) = \operatorname{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i \cdot K\left(\vec{\mathbf{x}}, \vec{\mathbf{x}}_i\right) + b\right)$$

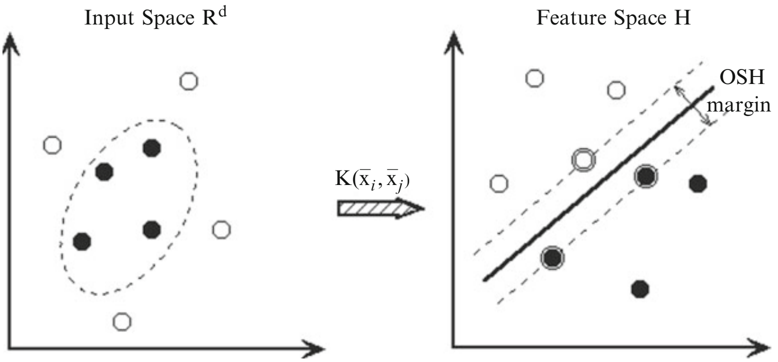


Fig. 10.24 Kernel function technology

The problem of risk minimization:

Given a set of functions

$$\left\{f\left(\vec{x},\alpha\right):\alpha\in\Lambda\right\},f\left(\vec{x},\alpha\right):\vec{x}\to\{-1,+1\},\vec{x}\in R^d$$

and a set of examples

$$\left(\vec{x}_i,y_i\right),\vec{x}_i\in R^d,y_i\in\{-1,+1\},i=1,2,\ldots,N$$

each one independently drawn from an unknown identical distribution.

The goal is to find an optimal function $f\left(\vec{x},\alpha^*\right)$ which minimizes the expected risk (or the actual risk) (Fig. 10.24).

$$R(\alpha)=\int L\left(f\left(\mathbf{\vec{x}},\alpha\right),y\right)\mathrm{d}P\left(\mathbf{\vec{x}},y\right)$$

$$\text{i.e. } R\left(\alpha^*\right)=\inf_{\alpha\in\Lambda}R(\alpha)$$

Here $L\left(f(\vec{x},\alpha^*),y\right)$ is the loss function. For this case one simple form is

$$L\left(f\left(\mathbf{\vec{x}},\alpha\right),\ y\right)=\frac{1}{2}\left|y-f\left(\mathbf{\vec{x}},\alpha\right)\right|,\mathbf{\vec{x}}\in R^d,y\in\{-1,+1\}$$

10.4.3.2 The Empirical Risk Minimization (ERM)

The risk functional $R(\alpha)$ is replaced by the so-called empirical risk function constructed on the basis of the training set:

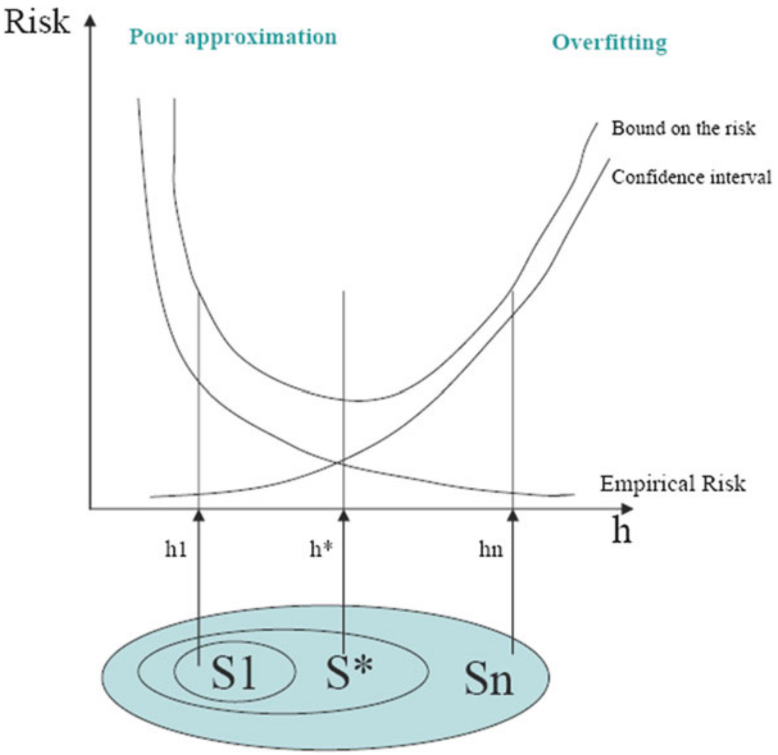


Fig. 10.25 SRM

$$R_{\text{emp}}\left(\alpha\right)=\frac{1}{N}\sum_{i=1}^{N}L\left(y_i,f\left(\vec{\mathbf{x}}_i,\alpha\right)\right)$$

$$R_{\text{emp}}\left(\alpha_e{}^*\right)=\inf_{\alpha\in\Lambda}\left\{R_{\text{emp}}\left(\alpha\right)\right\}$$

Note is $\alpha_e^*=\alpha^*$? No! The answer is not simple!

10.4.3.3 The Structural Risk Minimization (SRM)

The bound of generalization ability of learning machine (Vapnik & Chervonenkis):

$$R(\alpha)\leq R_{\text{emp}}\left(\alpha\right)+\Phi\left(\frac{N}{h}\right)$$

Here, N is the size of the training set; h, VC dimension, the measure of the capacity of the learning machine; and $\Phi(N/h)$, the confidence interval. When the N/h is larger, the confidence interval is smaller (Fig. 10.25).

10.4.3.4 New Approach to Protein Secondary Structure Prediction

The data sets

Two nonhomologous data sets:

1. The RS126 set – percentage identity – 25 %
2. The CB513 set – the SD (or *Z*) score – 5

We exclude entries if:

1. They are not determined by X-ray diffraction.
2. The program DSSP could not produce an output.
3. The protein had physical chain breaks.
4. They had a resolution worse than 0.19 nm.

10.4.3.5 Assignments of the Protein Secondary Structure

Now the automatic assignments of secondary structure to the experimentally determined 3D structure are usually performed by DSSP, STRIDE, or DEFINE.

Here we concentrate exclusively on the DSSP assignments, which distinguish eight secondary structure classes: H (α-helix), G (310-helix), I (π-helix), E (β-strand), B (isolated β-bridge), T (turn), S (bend), and (the rests).

We reduce the eight classes to three states – helix (H), sheet (E), and coil (C) according to two different methods:

1. DSSP: H, G, and I to H; E to E; and all other states to C
2. DSSP: H and G to H, E and B to E, and all other states to C

10.4.3.6 Assessment of Prediction Accuracy

Cross-validation trials are necessary to minimize variation in results caused by a particular choice of training or test sets.

A full jackknife test is not feasible, especially on the CB513 set for the limited computation power. We take the sevenfold cross-validation on both sets.

1. *Q*_index (Q_3, QH, QE,QC)
2. Matthews' Correlation Coefficient (CH, CE, CC)
3. Segment Overlap Measure (SOV)

*Q*_index gives percentage of residues predicted correctly as helix (H), strand (E), coil (C), or all three conformational states. The definition of *Q*_index is as follows:

1. For a single conformational state:

$$Q_{\mathrm{I}} = \frac{\text{Number of residues correctly predicted in state } i}{\text{Number of residues observed in state } i} * 100$$

where I is either H, E, or C.

2. For all three states:

$$Q_3 = \frac{\text{Number of residues correctly predicted}}{\text{Number of all residues}} * 100$$

10.4.3.7 The Coding Scheme

For the case of the single sequence, each residue is coded by the orthogonal binary vector (1,0,...,0) or (0,1,...,0). The vector is 21-dimensional. If the window length is l, the dimensionality of the feature vector (or the sample space) is 21*l. When we include the evolutionary information, for each residue the frequency of occurrence of each of the 20 amino acids at one position in the alignment is computed.

10.4.3.8 The Design of Binary Classifiers

We design six binary classifiers (SVMs) as follows:

1. Helix/non-helix – H/∼ H
2. Sheet/non-sheet – E/∼ E
3. Coil/non-coil – C/∼ C
4. Helix/sheet – H/E
5. Sheet/coil – E/C
6. Coil/helix – C/H

10.4.3.9 The Design of Tertiary Classifiers

Assembly of the binary classifiers:

1. SVM_MAX_D
   We combined the three one-versus-rest classifiers (H/∼H, E/∼E, and C/∼C) to handle the multiclass case. The class (H, E, or C) for a testing sample was assigned as that corresponding to the largest positive distance to the OSH.
2. SVM_TREE (Fig. 10.26)
3. SVM_NN (Fig. 10.27)

The tertiary classifiers we designed:

1. SVM_NN
2. SVM_TREE1
3. SVM_TREE3
4. SVM_TREE3
5. SVM_VOTE
6. SVM_MAX_D
7. SVM_JURY

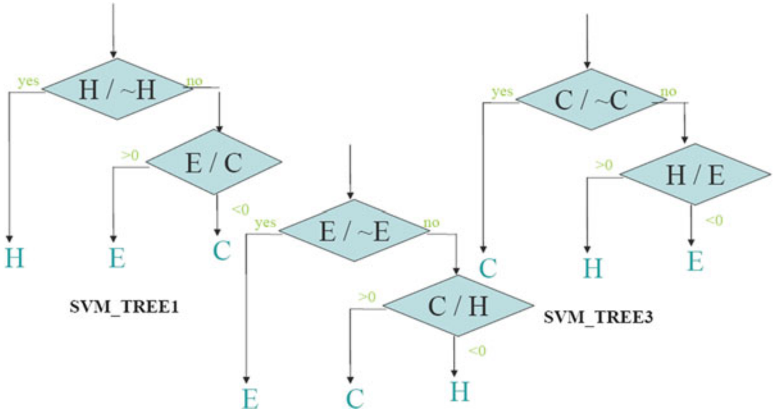
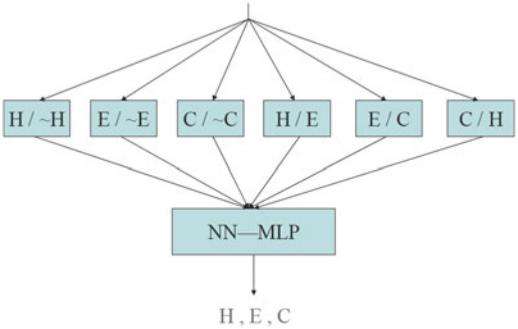


Fig. 10.26 SVM tree

Fig. 10.27 SVM-NN



10.4.3.10 Results and Analysis

The selection of the optimal kernel function and the parameters:

$$\text{RBF} \quad k(\vec{x}, \vec{y}) = \exp(-r|\vec{x} - \vec{y}|^2)$$

We set the optimal $\gamma = 0.10$.

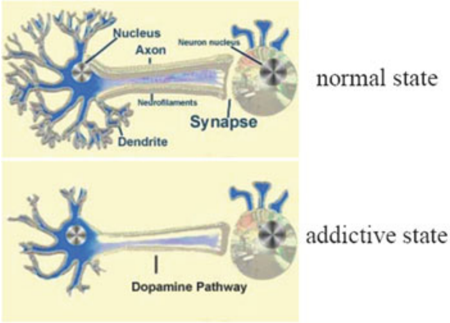
Accuracy measure

Three-state prediction accuracy: Q_3

$$Q_3 = \frac{\text{Correctly predicted residues}}{\text{Number of residues}}$$

A prediction of all loop: $Q_3 \sim 40\ \%$

Fig. 10.28 Neuron



Improvement of accuracy:

Chou and Fasman (1974)	∼50–53 %
Garnier (1978)	63 %
Zvelebil (1987)	66 %
Qian and Sejnowski (1988)	64.3 %
Rost and Sander (1993)	70.8–72.0 %
Frishman and Argos (1997)	<75 %
Cuff and Barton (1999)	72.9 %
Jones (1999)	76.5 %
Petersen et al. (2000)	77.9 %
Hua and Sun (2001)	76.2 %
Guo and Sun (2003)	80 %

10.4.3.11 Neural Network (Figs. 10.28, 10.29, 10.30, 10.31, 10.32, 10.33, and 10.34)

10.4.4 Homological Protein Structure Prediction Method

Homology modeling is a knowledge-based protein structure prediction. These kinds of methods are based on the evolutional conservation of protein structure and sequence. They use the structure of known proteins to build the structure of the unknown homological proteins. They are the most mature protein structure prediction methods so far. When the homology is high, we will get reliable prediction results. In the whole genome, only about 20–30 % sequences can be predicted using these methods.

One difficult point in the homology modeling method is the prediction of the circle region on the protein surface. That is because the circle region on the surface is very flexible. But because the circle region is usually the active part of the protein, the prediction of the structure of circle region is quite important to the protein structure modeling.

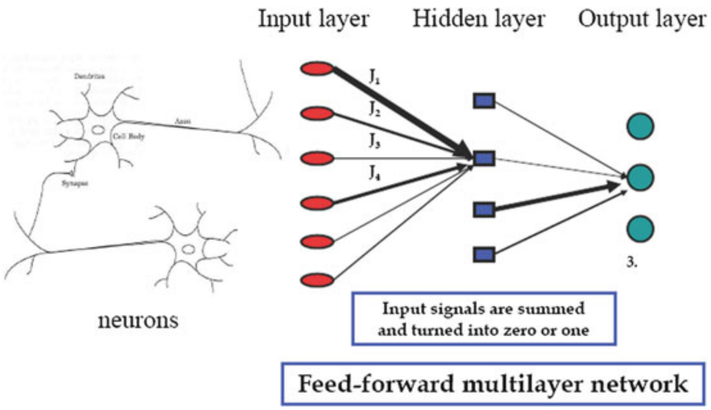
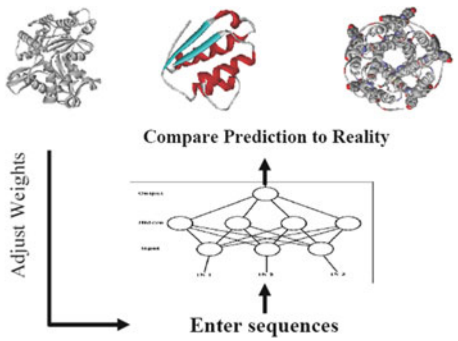

Input layer
Hidden layer
Output layer
Dendrites
Cell Body
Axon
Synapse
neurons
J1
J2
J3
J4
3.
Input signals are summed and turned into zero or one
Feed-forward multilayer network

Fig. 10.29 Neural network

Fig. 10.30 Neural network training



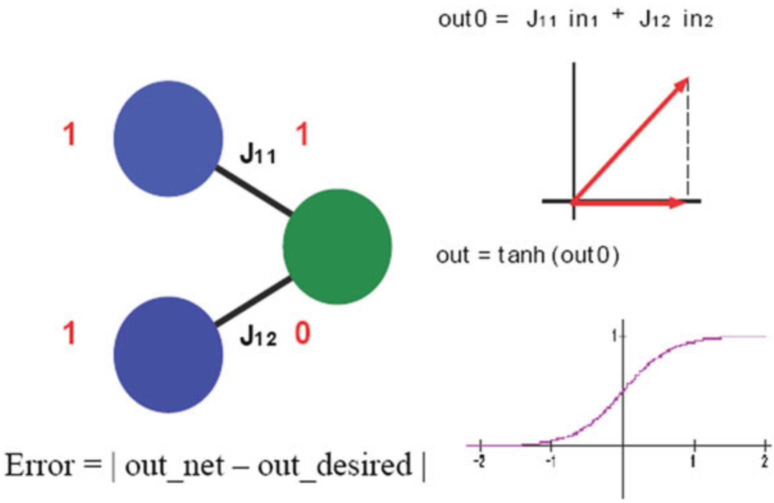


Fig. 10.31 Simple neural network

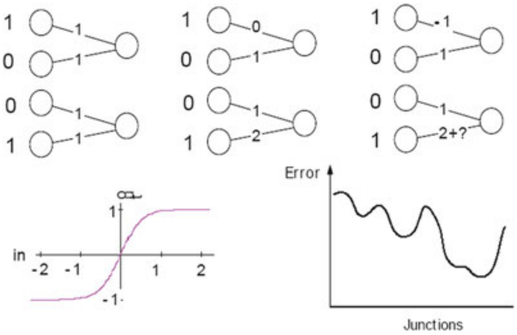
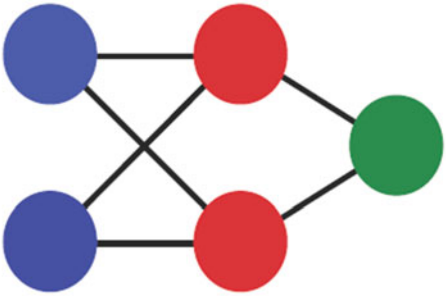


Fig. 10.32 Train a neural network



$$\text{out}_i = f\left(\sum_j J^2_{ij} \cdot f\left(\sum_k J^1_{jk} \cdot \text{in}_k\right)\right)$$

Fig. 10.33 Simple neural network with hidden layer

The protein homology modeling includes:

1. Matching of object protein sequence and model sequence
2. Modeling object protein structure model according to the model structure
3. Modeling the conserved region in the object protein
4. Modeling the SCRs backbone
5. Predicting the side chain structure
6. Optimizing and estimating the modeling structure

10.4.4.1 Threading Method

Threading (or inverse folding) method can be used to predict structure without homology information. The basic assumption is that the folding type of natural protein is limited. So we can align the sequence of proteins whose structures are unknown and those proteins whose structures are known. And then predict on the best alignment. This method cannot predict new types of proteins correctly.

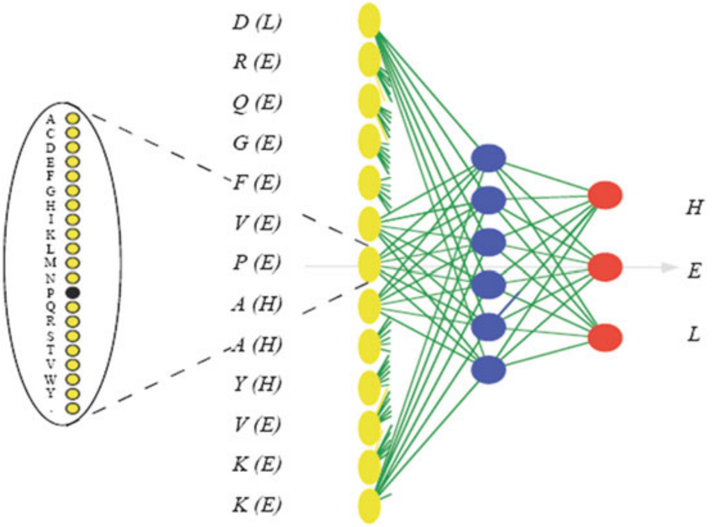


Fig. 10.34 Neural network for secondary structure

Threading method can be done by summarizing known independent protein structure patterns as the model of unknown structure and then by learning known database to summarize average potential function which can distinguish correction and error. In this way we can get the best alignment way.

Protein sequence incrustation:

1. Basing on the experience method. Build various potential functions by analyzing protein of known structure, and see if it can align with known structure by using the standard of lowest folding configuration to guide the object protein sequence incrustation.
2. Basing on the 3D profile. Predict sequence space structure by building a 3D profile, using dynamic programming, comparing new sequence with those in profile databases, and seeking optimal alignment.

10.4.5 Ab Initio Prediction Method

10.4.5.1 Protein Secondary Structure Prediction

Protein secondary structure prediction research has developed for more than three decades. From the progression of research method, there are three different periods. The first period is statistic prediction basing on single residue; the second period, statistic prediction basing on sequence segments; and the third period, statistic prediction combining evolutionary information.

Rost and Sander (1993) promoted prediction basing on neural network – PHD (Profile fed neural network systems from HeiDelberg). It is the first method with the prediction accuracy over 70 %, first efficient method bringing in evolutionary method, and one of the most accurate methods so far.

PHD is a complicated method basing on neural network. It includes poly-sequence information. Recently, Cuff and Barton synthesize many good secondary prediction methods, such as DSC, NNSSP, PREDATOR, and PHD. Up to now, there are still some other artificial intelligence methods to predict secondary structure, such as expert system and nearest neighbor method.

Recently, it is a good opportunity to predict protein secondary structure. For one thing, structural genomic plan is carried out throughout the world to increase the speed of measuring the number of protein structure and fold type. For another, the field of machine learning develops fast. For example, in recent 2 years, the building and perfecting of famous statistic learning theory of V. Vapnik make it possible for us to use the latest machine learning method to improve the prediction accuracy of secondary structure.

Our paper published on JMB (*J. Mol. Biol.*) applied SVM to predict protein secondary structure and got an accuracy of 76.2 %.

10.4.5.2 Chou-Fasman Method

1. Residue propensity factor

$$P_{ij} = \frac{f_{ij}}{f_j}$$

   j: configuration
   i: one of the twenty amino acids
   f_j: fraction of the jth configuration
   f_{ij}: jth configuration fraction of the ith amino acid residue. $f_{ij} = n_{ij}/N_i$
   n_{ij}: the total appearance of a residue in a certain configuration
   N_i: the total number of a residue in the statistical samples. $f_j = N_j/N_t$
   N_t: the total number of residues in the statistical samples

2. The tendentiousness of folding-type related secondary structure

   (a) Protein folding type: all α, all β, α + β, and α/β
   (b) Analysis of secondary structure tendentiousness: α-helix propensity factor P_α, β-sheet propensity factor P_β, and irregular curl propensity factor P_C

3. Chou-Fasman method

   (a) α-Helix rule
   In a protein sequence, there are at least four residues in the neighboring six residues tending to form α-helix kernel. The kernel extends laterally until the

average value of α-helix tendentiousness factor in the polypeptide segment $P_\beta < 1.0$. Lastly, drop three residues at each end of α-helix. If the rest part is longer than six residues, $P_\alpha > 1.03$, it will be predicted as helix.

(b) β-Sheet folding rule

If three residues in five tend to form β-sheet, we think it is the folding kernel. The kernel extends laterally until the average of the tendentiousness of the polypeptide segment $P_\beta < 1.0$. Lastly, discard two residues from each end; if the rest part is longer than the four residues and $P_\alpha > 1.05$, then it is predicted as β-sheet.

Reference

1. Vapnik VN (1995) The Nature of Statistical Learning Theory, Springer-Verlag