CHAPTER 1

HISTORICAL PERSPECTIVES

Biology becomes much more understandable in light of genetics (Ayala and Kiger 1984). This is true even more so in the case of the theory of evolution proposed by Darwin (1859). It seems the theory of evolution would have been placed on a solid foundation from the start if Darwin would have been aware of the Mendelian rules of inheritance. There is some indication that a copy of Mendel's publication was received by Darwin, which remained unopened during his lifetime. It is believed that this caused Darwin's failure to provide a firm basis on which selection works during the process of evolution.

Genetics has had several major breakthroughs during its development that have made biology a well-established discipline of science. Some of these break throughs are discussed here. The first major discovery was the rules of inheritance by Mendel (1866). This provided the particulate nature of inheritance and established the presence of genes, which control phenotypes. It also provided genes as the ultimate basis for propelling the process of evolution of organisms and integrated the different branches of the science of biology. In addition, Mendelian genetics transformed biology from a science based exclusively on observations to an experimental science where certain ideas could be tested by performing experiments.

The second major breakthrough was discovered by Beadle and Tatum (1941) with their conceptual one-gene–one-enzyme hypothesis. This proved the biochemical basis for the mechanism of gene action and integrated

Introduction to Proteomics: Principles and Applications, By Nawin C. Mishra

chemistry into biology. It provided the tool for analyzing metabolic pathways and several complex systems, including the nervous system. It also provided the understanding of the genetic basis of diseases and their possible cures by chemical manipulations and ultimately by gene therapy.

The discovery of the structure of DNA by Watson and Crick (1953) marked the third major breakthrough in biology. The discovery of the Watson–Crick DNA structure was aptly meaningful in view of the findings of DNA as the chemical basis of inheritance (Avery et al. 1944, Hershey and Chase 1952). The Watson–Crick structure of DNA provided the molecular basis for the understanding of the mechanisms of the storage and transmission of genetic information and possible changes (mutations) therein. Mutation provided the source of variations that could be selected for during the process of Darwinian evolution. Thus, the DNA structure created by Watson and Crick made genetics not only necessary but also unavoidable in the understanding of Darwin's evolution by natural selection. In 1962, Watson, Crick, and Wilkins received the Nobel Prize for this landmark discovery of the DNA structure.

The development of the Watson–Crick structure of DNA led to the birth of molecular biology followed by the enunciation of the central dogma in biology. Molecular biology attempted to provide the molecular basis for everything in biology and biochemistry leading to the unity of life. Molecular biology perpetuated the reductionistic view of living systems: Reductionists attempt to understand a system by understanding its molecular components. Molecular biology also led to the development of a better understanding of diseases and their control by pharmaceuticals. The field of molecular biology ushered in by the Watson–Crick DNA structure led to the development of scores of Nobel Prize-winning concepts in biology, biochemistry, and medicine as discussed later in this book.

The coming of genomics marked the fourth major breakthrough in biology. Advances in genome sequencing and availability of human and several other genome sequences by 2001 provided the basis for the understanding of the uniqueness of humans in possessing certain distinctive DNA segments. Genomics also provides the basis for the understanding of variations among individuals as differences in DNA sequences. Furthermore, it provides molecular insight into the genetic basis for differences in our response to the same drug. The variation in individual DNA sequences is expected to provide the molecular understanding of our several complex traits, including behavior. DNA sequences also provide a better insight into the record of the evolutionary processes in an organism. Genomics is expected to provide a better understanding of a complex organism like humans after the elucidation of the roles of noncoding sequences (introns) of DNA. Understanding the roles of introns is currently a formidable task: It is believed

that the elucidation of the roles of introns will add a new dimension to the understanding of biology.

The fifth breakthrough underway is the development of proteomics. This is bringing a better understanding of biochemical pathways and the roles of protein interactions. Above all, proteomics provides a clue to answering the big question of how a small number of genes can control several phenotypes in a complex organism like humans. A major conceptual scheme emerging from proteomics is that it is the number of interactions of proteins and not the number of proteins per se that is responsible for the myriad phenotypes in an organism.

The sixth breakthrough that is in making involves the science of synthetic genetics which would allow creation of new organisms by creation of entirely new genomes or by the manipulation of existing ones with the help of the techniques of molecular genetics, genomics, proteomics and bioinformatics.

Advances in genomics and proteomics in conjunction with bioinformatics have made it possible to realize the dreams of the chemists of the 20th century. These chemists wanted to decipher the amino acid sequences of all proteins to understand their functions. Proteomics has made it possible to determine the amino acid sequence of any protein. In addition, future advances in genomics and proteomics are expected to bring several revolutions in medicine and will make personalized medicine a reality. Advances in proteomics are expected to integrate the reductionistic views of Watson and Crick into systems biology to show how molecular parts evolved and how they fit together to work as an organism. The latter is expected to provide the ultimate understanding of biology.

1.1 INTRODUCTION TO PROTEOMICS

The term "proteome" originates from the words protein and genome. It represents the entire collection of proteins encoded by the genome in an organism. Proteomics, therefore, is defined as the total protein content of a cell or that of an organism. Proteomics is the understanding of the structure, function, and interactions of the entire protein content of an organism. Proteins control the phenotype of a cell by determining its structure and, above all, by carrying out all functions in a cell. Defective proteins are the major causes of diseases and thus serve as useful indicators for the diagnosis of a particular disease. In addition, proteins are the primary targets of most drugs and thus are the main basis for the development of new drugs. Therefore, the study of proteomics is important for understanding their role

in the cause and control of diseases and in the development of humans as well as that of other organisms.

Proteins are encoded by DNA in most organisms and by RNA in some viruses. In all cases except RNA viruses, DNA is transcribed into RNA, which is then translated into a protein. In case of RNA virus, however, RNA is translated directly into proteins. Initially, it was thought that one gene makes one enzyme, which controls a phenotype. However, this view has undergone tremendous changes in the last several decades mainly because of the discovery of the split nature of eukaryotic genes, which involves RNA splicing, the occurrence of RNA editing, and the phenomenon of RNA silencing. The split nature of gene, RNA splicing, RNA editing, and RNA silencing are discussed later in this chapter.

In eukaryotes, the coding sequences of a gene called exons are interrupted by the noncoding stretches of nucleotides called introns. The exons are spliced after removal of introns within a gene continuously (referred to as cis splicing) or discontinuously (referred to as alternate splicing) or between exons of different genes leading to transsplicing. The different modes of splicing of exons and posttranslational modifications of proteins are responsible for the abundance of proteins in eukaryotic organisms. In humans there are approximately 23,000 genes and more than 500,000 proteins.

The findings of suppressor genes and the split nature of genes may present apparent contradictions to the one-gene–one-enzyme hypothesis. However, with the coming of central dogma (Crick, 1958, 1970, Watson 1965, Mattick 2003, Lewin 2004) in biology and elucidation of the genetic code (Leder and Nirenberg 1964, Khorana 1968), it is understandable how suppressor genes work. Thus, the mechanism of action of suppressor genes does not contradict the original ideas implicit in Beadle and Tatum's one-gene–one-enzyme concept to any extent as it appears superficially. In light of central dogma, it is understandable that certain genes or DNA segments may code for different proteins or that the coding section of protein in DNA is distributed across a huge expanse of DNA interrupted by the noncoding sequences. It has become obvious that the one-gene–one-enzyme concept applies only to genes that encode one polypeptide and not to genes that have a split nature and can code more than one protein. Thus, the one-gene–one-enzyme concept is limited to the nature of the gene itself, just as Mendelian rules of inheritance apply only to the genes located in the nucleus and not to the genes that are located elsewhere in the cell beyond the nucleus. Thus, the Mendelian inheritance pertains to the location of the genes, whereas the one-gene–one-enzyme concept is limited to the nature of the gene itself.

Obviously, what Beadle and Tatum suggested is not an axiom but a rule, and certain situations just represent exceptions to their profound rule. It

seems that nature too has the British view of rule that "exceptions prove the rule." The history of science is full of such exceptions. The most glaring example of such an exception involves the central dogma in molecular biology described by Francis Crick, the codiscoverer of the DNA structure. Crick (1958, 1970) surmised that sequential information in DNA is transferred to RNA and then to protein from RNA and that the direction of this information transfer is fixed. However, later it was shown that RNA is reverse transcribed into DNA, and at times, messenger RNA (mRNA) is edited by the addition or removal of cytidine or uridine before its translation in to protein, which suggests that information in a DNA segment is not translated directly into protein as implicit in central dogma. This idea suggests that DNA makes RNA, which makes protein. Howard Temin and David Baltimore received the Nobel Prize in 1975 for demonstrating this reverse transfer of information from RNA to DNA. The other glaring example of such an exception includes the enzymes. It was James Sumner of the Cornell University who established that enzymes are proteins. Soon, enzymes became synonymous with proteins until Sydney Altman of Yale University and Thomas Cech of the University of Colorado showed independently that certain enzymes are made of RNA and not proteins. Sumner in 1946 and Altman and Cech in 1989 were awarded Nobel Prizes for their contributions to the science of chemistry. Thus, it seems that biology, like any other branch of science, is replete with instances of exceptions to the rules.

The Swedish scientist Berzelius (1838)[1] named certain naturally occurring polymers as proteins. The fact that enzymes are proteins was established by Sumner (1946). Later, Sanger (1958)established that proteins are made up of a sequence of amino acids. The fact that an enzyme and a substrate (or an antibody and antigen) require precise complementary fit in their structures, just like a hand in a glove, to interact with each other was established by Linus Pauling in the 1940s. In addition to Sumner (1946), both Pauling (1954) and Sanger (1958) received Nobel Prizes for their work in chemistry. Most proteins have enzymatic functions, but several of them such as actin and fibrinoactin are structural components of cells. Proteins are major constituents of muscle, cartilage, and bones. Proteins are also responsible for the mobility of muscle cells. Certain proteins serve as receptors for different molecules or work as immunoglobulins or antigens, or proteins can serve as allergens or participate in transport of various molecules, such as oxygen or sex hormones. Many proteins are hormones, such as insulin or human growth hormone (HGH), which control important

[1]The word protein was coined from the Greek word proteios first by Jöns Jakob Berzelius in 1838 in a letter to his friend.

metabolic functions in humans and other organisms. The three-dimensional structure and chemical modifications of proteins are important for the understanding of their functions in different capacities.

Gorrod (1909) first described certain human disorders as inborn errors of metabolism and implied the genetic basis of these diseases. However, it was the genius of Beadle and Tatum (1941) that led to the establishment of the fact that a protein is encoded by a gene. Working with, Neurospora, they showed that the synthesis of a substance in a metabolic pathway was impaired in a mutant. They showed that by disabling the gene controlling the enzyme that catalyzed a biochemical reaction in a metabolic pathway, the mutant developed nutritional requirements for that substance. Such mutants could not be grown on a minimal medium, but their growth was possible only when a particular substance was added to the minimal medium. For example, a mutant with impaired synthesis of arginine could not be grown on a minimal medium, but its growth was possible only when arginine was added to the minimal medium. This method was also used to map the biochemical pathways.

Beadle and Tatum (1941) called this conceptual scheme the one-gene–one-enzyme hypothesis. This hypothesis has been modified in various ways. However, despite several exceptions to this rule of one gene encoding one enzyme, the main tenets of the one-gene–one-enzyme hypothesis have remained the cornerstone of biology. This concept has been instrumental for the merger of chemistry with genetics and for the development of molecular biology. This theory provides the standard method to assign a function to a protein by creating a mutant and then showing which protein has a defective function or which function has been impaired in a particular protein. Because of this hypothesis, it was possible to analyze and study viral, microbial, plant, and animal genetics. This has been the basis for creating knockout mutations and for in vitro mutagenesis. This hypothesis has proven crucial for the analysis of any basic genetic mechanism, such as DNA replication, repair, and recombination, and for establishing the role of a protein in any metabolic pathway. Finally, this theory by Beadle and Tatum has led to advances in agriculture, animal husbandry, pharmaceutical sciences, and medicine. The one-gene–one-enzyme hypothesis has been the basis for the understanding and alleviation of human diseases and for the development of gene therapy.

The one-gene–one-enzyme hypothesis implied that a mutant must have altered the protein. Beadle and Tatum could not demonstrate the defective nature of the protein in their mutants because of the lack of technology at that time. However, this was demonstrated first at the biochemical level by Mitchell and Lein (1948, Mitchell, et al. 1948) and by Yanofsky (1952, 2005a,b) in tryptophan, which required mutants of Neurospora that lacked

the enzyme tryptophan synthetase responsible for the synthesis of tryptophan. This concept was also demonstrated later at the molecular level by Ingram (1957) in the case of hemoglobin in persons who suffer from sickle cell anemia. Ingram showed that the sixth amino acid "glutamic acid," which is found in the hemoglobin of a normal person, is replaced by valine in the hemoglobin of a sickle cell person. This one change from glutamic acid to valine is the basis for the blood disorders in a sickle cell person. Later, many other mutants were shown to lack a protein altogether or possess proteins with altered amino acid(s).

The one-gene–one-enzyme theory also implied the correspondence in the ordered position of nucleotides in a gene with the position of amino acid in the protein encoded by that gene. This colinearity in the structure of a gene and that of a protein was demonstrated independently by Yanofsky et al. (1964) and by Sarabhai, et al. (1964), as discussed later in this chapter.

1.2 PROTEOME AND PROTEOMICS

1.2.1 Proteins as the Cell's Way of Accomplishing Specific Functions

The proteome is defined as the total proteins encoded by the genome of an organism. Proteomics is the science of describing the identification and features of the proteome of an organism.

The term "proteome" was first used by Marc Wilkins in 1994 (Wilkins 1996). An effort to describe the total proteins of an organism was made independently by O'Farrell (1975) and by Klose (1975). They developed what is called two-dimensional (2D) gel electrophoresis by running gel electrophoresis of proteins in two planes at right angles to each other (O'Farrell 1975, Klose 1975). This method separated a complex mixture of more than 1100 proteins of *Escherichia coli* into distinct bands of individual components on the gel. Later, the science of proteomics was revolutionized by the application of mass spectrometry in conjunction with genomics for the separation and identification of proteins on a large scale.

The genome of an organism is static in the sense that it remains the same in all cell types all the time. In contrast, the proteome of an organism is dynamic, because it differs from one cell type to another and keeps changing even in the same cell type at the different stages of activity or different states of development. A change in the proteome is a reflection of differential activity of the genes dependent on the cell type to express the protein needed for a particular function. For example, blood cells predominantly express the hemoglobin gene to produce the hemoglobin protein required

for the transport of oxygen, whereas pancreatic cells largely express the insulin gene, which produces the insulin peptide required for the entry of glucose molecules into cells.

Thus, the differential expression of genes is required for the production of different proteins because each protein controls a distinct function. The function of many proteins is listed in Table 1.1. In addition, the protein profile of a cell can vary depending on the different kinds of modification of the same protein; such modifications of protein may involve acetylation, phosphorylation, glycosylation, or association with lipid or carbohydrate molecules. These modifications in proteins occur as posttranslational events and alter the function of proteins. One example is the mitosis activator protein (MAP) kinase protein controlling the mitosis; this protein is activated by phosphorylation to give MAP Kinase (MAPK), MAP kinase kinase (MAPKK), and MAP kinase kinase kinase (MAPKKK). The role of protein modification in the control of cellular activity is discussed later in this book.

1.2.2 Pregenomic Proteomics

The role of proteins as enzymes in controlling a cellular activity was known much before its structure was elucidated. The conceptual breakthrough in deciphering the structure of a protein as a linear array of amino acids came from the enunciation of the one-gene enzyme concept. This conceptual breakthrough was materialized by certain technical advances. The technical advances included the development of machines for the analysis of the amino acid composition and for the determination of the sequence of the amino acids in a protein. With the help of these machines, the structure of proteins was elucidated one protein at a time for several years. Later,

Table 1.1. Function of different proteins.

Function	Protein
1. Catalyst	Enzymes (more than 90% of proteins) Catalyze biochemical reactions in the cell
2. Transport	Hemoglobin (carrier of oxygen) Albumin (carrier of hormones)
3. Structure	Cartilage/bone proteins
4. Cellular skeleton	Actin, fibrinoactin
5. Hormone	Insulin, growth hormone
6. Antibody	Immunoglobulins
7. Antigens and allergens	Bacterial and viral proteins
8. Mobility/muscle movement	Myosin
9. Receptors	Receptor for cholesterol
10. Cell communication/signaling	Transduction proteins, junction proteins

the introduction of the methodology of the 2D gel and that of mass spectrometry facilitated the simultaneous resolution of the structure of several proteins at the same time. Understanding the structure of several proteins at the same time aided by mass spectrometry was moved forward with the coming of genomics and bioinformatics. The methods of genomics deciphered the nucleotide sequence of DNA/genes in the chromosomes of various organisms. The methods of bioinformatics involved the use of computers and several software programs for analyzing the bulk of the nucleotide sequence of DNA of an organism. Bioinformatics is also used for deciphering the amino acid sequence of a protein from the sequence of nucleotides in a DNA molecule.

1.3 GENETICS OF PROTEINS

A genetic approach to understanding protein structure and function was dictated by the one-gene–one-enzyme hypothesis. This concept implied that the structure and function of proteins could be understood by the comparison of the protein obtained from the wild type and from mutant organisms. In reality, it became a routine method to understand the role of a protein in any metabolic or developmental pathway. Following this dictum, the hemoglobin molecules from normal humans and from sickle cell patients were compared. The hemoglobin of normal individuals was found to be different from the sickle cell patients in the sixth amino acid. Normal individuals possessed glutamic acid at this position, whereas the sickle cell patient possessed valine (Ingram 1956, 1957). Thus, one change in amino acid completely altered the structure and metabolic role of hemoglobin (Figure 1.1).

1.3.1 One-Gene — One-Enzyme Theory

This theory proposed by Beadle and Tatum (1941) implied that the structure of an enzyme or a protein is controlled by one gene, in the sense that one gene encodes one protein. This theory became useful in understanding

	1	2	3	4	5	6	7	8
Hemoglobin A	Val–	His–	Leu–	Thr–	Pro–	*Glu*–	*Glu*–	*Lys*–
Hemoglobin S	Val–	His–	Leu–	Thr–	Pro–	*Val*–	Glu–	Lys–

Figure 1.1: A comparison of the N-terminal amino acid sequence in the beta chain of hemoglobin of normal and sickle cell patients.

the biochemistry of any metabolic pathway and the role of proteins that catalyzed the biochemical reaction at each step in that metabolic pathway. First, it became obvious that if an organism cannot grow without a supplement, such as a specific amino acid, nucleotide, or vitamin, then that organism is defective for the protein that catalyzes the biochemical reaction leading to the synthesis of that substance, which has become a nutritional requirement for its growth.

This led to the development of a methodology to identify mutants with a specific nutritional requirement and then the order of biochemical reactions in a metabolic pathway. Such an analysis of nutritional mutants revealed the presence of a different class of mutants. Among them, a class of mutants was found to require the amino acid ornithine or citrulline, or arginine for growth. Another group of mutants required either citrulline or arginine for growth, whereas the third group of mutants could grow only in the presence of arginine. The nutritional requirement of this last group of mutants was not met by adding ornithine or citrulline as a supplement to the growth medium when added alone or together. The nutritional requirements of these three groups of mutants suggested a metabolic pathway for the synthesis of arginine by the organism. Thus, this metabolic pathway involved the sequential steps of biochemical reactions involving the synthesis of ornithine from a precursor molecule and then the synthesis of citrulline from ornithine, and finally arginine from citrulline. Therefore, the metabolic pathway was established as follows: Precursor $\rightarrow$ Ornithine $\rightarrow$ Citrulline $\rightarrow$ Arginine. From this sequence of biochemical reactions in this pathway, it becomes obvious that the first group of mutants is defective in the step involving the conversion of the precursor into ornithine. Therefore, this group of mutants could use either ornithine, citrulline, or arginine for growth. The second group of mutants is defective in the step involving the conversion of ornithine into citrulline; therefore, its growth requirement could be satisfied by the addition of citrulline or argine but not ornithine. The third group of mutants is defective in the last step of biochemical reaction involving the conversion of citrulline into arginine, and thus, an organism could grow only when arginine is added as the supplement. Thus, the one-gene–one-enzyme concept became a useful tool in establishing the sequence of biochemical reactions in a particular pathway. This theory also implied that if the enzyme catalyzing the conversion of substance A into substance B is defective, then the molecules of substance A will accumulate in the organism. At times, the accumulation of this substance may cause a hazard to the health of mutant individuals. This is shown by the accumulation of phenylalanine in phenylketoneurics or the accumulation of homogentisic acid in infants who suffer from alcaptonuria. Such metabolic

blockages occur in the metabolic pathway of phenylalanine–tyrosine pathways as a result of the specific enzyme defects, as observed in Figure 1.2. Such genetic defects were described as "inborn errors of metabolism" by Gorrod (1909). An accumulation of phenylalanine causes damage to the development of the brain in early stages of development, and it could lead to mental retardation. Now it is mandatory in the United States and other developed countries to screen babies after birth to check for phenylketoneuria by evaluating for an increased amount of phenylalanine in the blood. Phenylketoneuric babies are put on a special diet deficient in protein to manage the level of phenylalanine. After brain development is complete, these individuals are returned to a normal diet. However, a phenylketoneuric female must restrict the phenylalanine intake during pregnancy to allow the proper growth development of the infant's brain.

Later, this theory became useful in establishing the identification of a particular protein and its role in a biochemical step in the metabolic pathway by

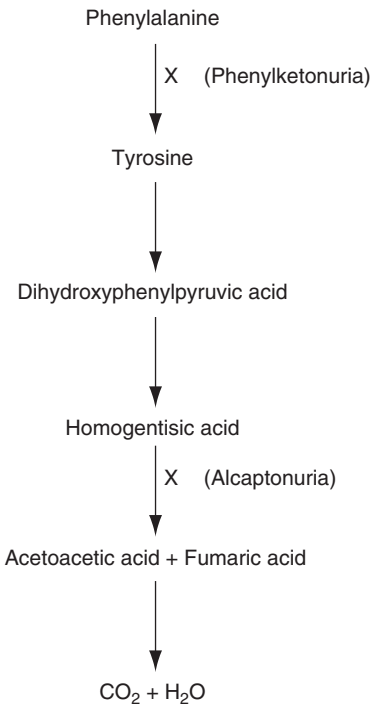

Figure 1.2: Consequences of a metabolic block in pheylalanine–tyrosine Defective phenylalanine hydroxylase can lead to the accumulation of phenylalanine, which can cause damage to brain cells and mental retardation in phenylketonuric babies. Another metabolic blockage caused by a defective enzyme can lead to alcaptonuria.

comparing the biophysical properties of the wild-type and mutant enzyme involved in the particular pathway. It was soon found that a mutant did not produce a particular protein, or produced a partial protein, or a defective protein with a different amino acid in a certain position in the protein. The occurrence of distinct classes of mutant proteins is consistent with the nature of changes that accompany a change in the genetic code. Such a change may involve the substitution of one nucleotide by another in the genetic code or a deletion or insertion of a nucleotide in the DNA sequence of the gene. A substitution of nucleotide in the genetic code may cause a nonsense, missense, or silent mutation in the protein. A nonsense mutation results from a change in the existing amino acid codon into a stop codon. A nonsense mutation that occurs in the beginning of a gene encoding the protein will make a small peptide or no protein at all. A nonsense mutation anywhere in the gene will yield a truncated protein of different lengths. A missense mutation that causes the substitution of one amino acid for another amino acid may alter the biochemical properties of the protein so that it is rendered inactive or partially active. However, such a substitution of one nucleotide by another in the genetic code may not cause any change in the resulting protein because of degeneracy of the genetic code or because a replaced amino acid may have no adverse effect on the overall structure and function of the protein. Such mutations are called neutral or silent mutations. A deletion or insertion of a nucleotide in the genetic code leads to a shift in the reading of the triplet genetic code. Such a frame shift mutation leads to changes in the nature of all amino acids from the point of insertion or deletion of the nucleotide. If it occurs in the beginning or middle of the gene, then it causes changes in a large number of the amino acids in the resulting protein, rendering that protein completely inactive. However, if the insertion or deletion of a nucleotide occurs toward the end of the gene, it is possible that the resulting amino acid changes may still leave the activity of the protein intact. All these kinds of mutations have been found to occur in the genome of an organism.

One-gene–one-enzyme theory suggested that a mutant would lack a protein or possess a defective protein. This was shown first in tryptophan requiring a Neurospora mutant and then later in similar mutants of *E. coli*. Currently, hundreds of mutants have been analyzed, which shows this one-to-one relationship in gene and protein with mutants always possessing no protein or a defective protein that lacks enzyme activity. Thus, one-gene–one-enzyme theory provided not only the informational role of the gene in encoding a protein but also provided a tool to dissect the biochemistry of any simple to complex processes in the living system by producing mutants and then comparing the biochemical changes in the mutant. No system has escaped the scope of this powerful tool.

1.3.1.1 Colinearity of Gene and Protein. The one-gene–one-enzyme concept of Beadle and Tatum (1941) provided the basis for colinearity in the DNA/gene and protein structures with a suggestion that the gene represents a sequence of nucleotides and the protein represents a sequence of amino acids. Avery et al. (1944) and Hershey and Chase (1952), by their transfection experiments in bacteria and bacterial viruses, established that genes are made up of DNA molecules. The fact that the gene is a sequence of nucleotides was shown by the correspondence between the genetic map of certain mutants with blocks of nucleotides. This colinearity between the DNA sequence of genes and the amino acid sequence of proteins was established by the study of missense mutants of *E. coli* (Yanofsky et al. 1964) or of nonsense mutants of a bacterial virus (Sarabhai et al. 1964). In both cases, the position of change in the genetic code corresponded with the position of amino acid change in the protein. Yanofsky et al. showed that a change in the early nucleotide sequence of a bacterial gene for protein A of tryptophan synthetase caused a corresponding change in the early amino acids in the protein. A change in the middle of the gene corresponded with a change in amino acid position in the middle of the protein. Similarly, a change in the end of a gene corresponded with a change in position toward the end of protein A of tryptophan synthetase. Sarabhai et al. (1964) showed that a virus produced truncated viral proteins; the size of the peptides corresponded with the length of the gene where the nonsense mutation occurred (Figure 1.3).

1.3.1.2 Protein as a Sequence of Amino Acids. The fact that a protein is a sequence an amino acid was directly established by the elucidation of the structure of insulin polypeptide as a linear sequence of different amino acids by Sanger (1958). Thus, insulin was the polypeptide or a small protein that was sequenced first by Sanger (1958). Ribonuclease A was the first full-size protein and an enzyme that was sequenced by

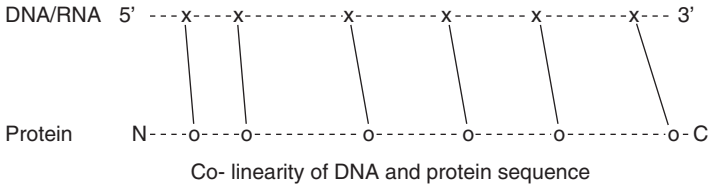

Figure 1.3: Colinearity of the DNA and protein sequence. The X represents the site of mutation in the gene/DNA as mapped by recombinational analyses. The O represents the position of altered amino acids in the protein coded by the gene. Vertical lines connect the position of changes in the gene and protein to show their one-to-one correspondence.

Stein and Moore (1972). However, the direct demonstration that a gene is a sequence of nucleotides was accomplished much later when the method for cloning of a gene and its sequence analysis became available. Proteins usually have four kinds of structure before a three-dimensional structure is assumed. These different structures are called a primary, secondary, tertiary, and quaternary structure (Figure 1.4). The linear sequence of amino acids in the proteins represents the primary structure. The secondary and tertiary structures originate from the folding of polypeptide on itself as a result of the interaction of the side groups attached to the amino acids. The quaternary structure results from the interaction of two or more fully folded polypeptides that interact with each other to give the protein structure.

The one-gene–one-enzyme concept did imply that the primary structure of the peptide determines the secondary, tertiary, and quaternary structure, and this was established by Anfinsen (1973) by an analysis of the mutant ribonuclease and by the study of chemical modification as well as the denaturation and renaturation kinetics of this enzyme (Anfinsen 1973).

1.3.1.3 One Gene—Many Proteins: Challenge to Proteomics. The central dogma of biology suggests the direction of the flow of genetic information from DNA to RNA to protein is DNA → RNA → Protein.

In this scheme, the one-gene–one-enzyme concept of Beadle and Tatum is written as follows: One DNA → One RNA (Transcript or mRNA) → One protein This scheme holds well for the prokaryotic organisms, because in prokaryotic genes, the protein-encoding information is continuous and the transcript is directly translatable and equivalent to mRNA. However, it was soon found that many genes in eukaryotes have a split gene structure in that the protein-encoding segments (exon) in a gene may be interrupted by noncoding segments (intron). In view of the split nature of many eukaryotic genes, the transcript must undergo a process to remove the noncoding intervening sequences (introns) to make all coding segments or exons continuous to yield mRNA, which is translatable. The splicing of exons may occur in different ways and can lead to different kinds of mRNA from the same transcript.

Thus, because of the split nature of the eukaryotic genes, the Beadle and Tatum concept of gene–enzyme relation has to be modified, as one gene can create many proteins and could be written in the language of central dogma as

$$\text{One DNA} \rightarrow \text{One transcript} \rightarrow \text{many mRNAs} \rightarrow \text{Many proteins}$$

It is of interest to note that the central dogma changed when it was found that RNA could be reverse transcribed into DNA. The central dogma is

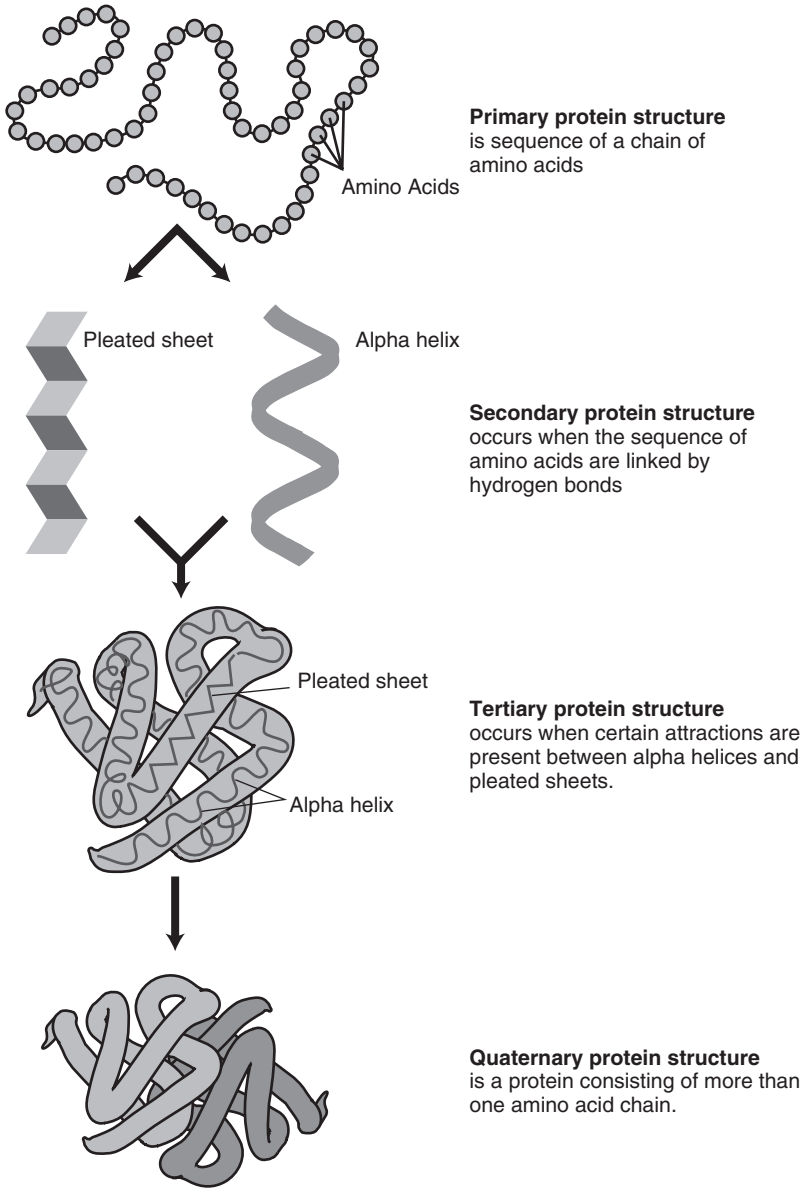

Figure 1.4: Structure of protein with different levels of organization. Reproduced with permission of Darryl Leza of NIHGR/NIH.)

now depicted as

$$\text{DNA} \leftrightarrow \text{RNA} \rightarrow \text{Protein, instead of} \text{DNA} \rightarrow \text{RNA} \rightarrow \text{Protein}$$

Thus, the central dogma is no more an axiom and that is true of Beadle and Tatum's one-gene–one-enzyme concept as well. Indeed they represent certain profound rules in biology. However, these rules have to be modified to accommodate new facts regarding the nature of gene as new facts emerge.

The new idea that one gene may encode many proteins has helped in understanding how only 23,000 genes in the human can code for more than 90,000 proteins. In the pregenomic era, it was thought that humans may have 100,000 genes or more. However, the results of the human genome project revealed the presence of approximately 23,000 protein-encoding genes; this paradox is resolved by the dictum that one gene makes one transcript, but one transcript gives rise to many mRNAs, which are in turn translated into many distinct proteins. Thus, it is possible that more than 90,000 proteins in humans can be encoded by 23,000 human genes. In many higher eukaryotes such as primates (including humans) and in rodents, more than 50% of genes code for more than one protein (Lander et al. 2001). In Drosophila, it has been estimated that a particular gene DSCAM encodes more than 38,000 proteins. The number of proteins in the different human cells at different stages is estimated to be approximately 500,000; this increase in the number of proteins in human cells results from posttranslational modifications of the 90,000 proteins encoded by 23,000 human genes. Finally, it is pertinent to point out that in prokaryotes, almost 100% of genes encode one protein per gene.

In lower eukaryotes such as yeast or filamentous fungi, only approximately 90% of genes encode one protein per gene. This picture changes dramatically in higher organisms including humans, where more than 50% of genes encode one protein per gene, whereas other genes encode more than one protein per gene. It seems that on average, one gene codes for more than three proteins in higher eukaryotes.

1.3.2 RNA Splicing

In higher organisms, a gene is first transcribed into a transcript or pre-mRNA. The latter undergoes additional modifications called "processing" to produce translatable mRNA. The processing involves at least three steps. The first step includes a cap or the addition of novel guanosine nucleotide at the 5'end, and the second step includes a tail or the addition of a poly A nucleotides at the 3'end. The third step is the removal of

intervening noncoding sequences called introns from the transcript. RNA splicing accomplishes the removal of introns and the joining of exons so that the different coding sequences in a transcript become continuous in the resulting mRNA. RNA splicing is carried out by a complex of RNAs and proteins organized into an organelle called a splicosome. A splicosome is as big as a ribosome and provides the platform on the surface of which the joining of exons and removal of introns are carried out. The two ends of an intron are recognized by certain concensus sequences such as GA at the 5'end and GU at the 3'end of the intron. During the process of RNA splicing, an intron loops out and is removed as a lariate structure with a guanine nucleotide as the tail bringing the neighboring exons together. Some introns are self-splicing and are removed without a splicosome. The RNA splicing of pre-mRNA occurs exclusively in eukaryotes. However, certain transfer RNAs (tRNAs) may undergo splicing in both prokaryotes and eukaryotes; their splicing is carried by out by certain enzymes without the involvement of splicosomes.

Eukaryotic pre-mRNA may be spliced out in different ways. First, the different exons of a particular pre-mRNA are brought together continuously by the removal of introns, which yields one translatable mRNA. For example, a pre-mRNA containing three exons and two introns will produce a mRNA after the removal of intons with all three exons together; such mRNA will produce a long protein on translation. Second, the different exons of this or similar pre-mRNAs may undergo alternate splicing, which yields several translatable mRNAs. For example, a pre-mRNA with three exons and two introns may undergo alternate splicing, which produces two different messages, one mRNA with exon one and exon two together, and other mRNA with exon one and exon three together. Thus, these two mRNAs will produce different proteins during translation. At times, certain exons of two different pre-mRNAs may be spliced together to yield different mRNAs. Such splicing that involves the exons of different pre-mRNAs is called transsplicing (Figure 1.5).

The process of alternate splicing is the major cause for the production of many proteins from one gene. The process of transsplicing causes the formation of one or more proteins from two genes. These two situations represent a major departure from the original one-gene–one-enzyme theory of Beadle and Tatum (1941). However, at the molecular level, it seems logical because enzymes or proteins are made up of modules encoded by the exons. Thus, nature has evolved ways such as alternate splicing and transsplicing to bring these modules together to produce a functional enzyme or protein.

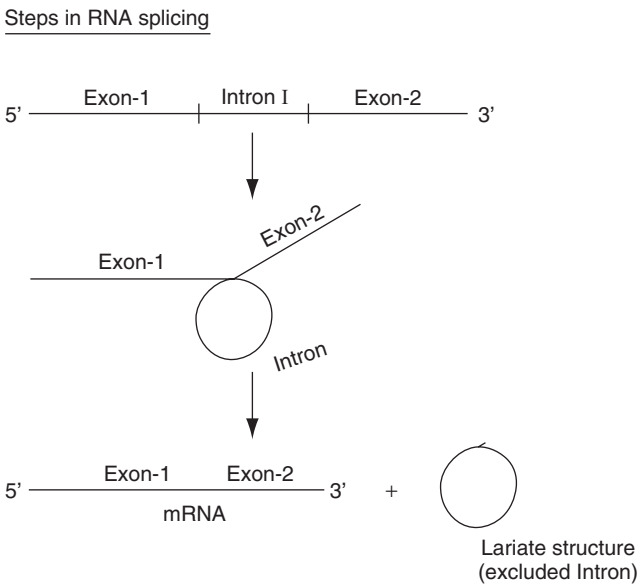

Figure 1.5: Removal of intron from a transcript.

1.3.3 RNA Editing

In addition to RNA splicing, the process of RNA editing is another factor that changes the nature of proteins. One gene may produce more than one functional protein through RNA editing. Thus, RNA editing can influence the proteomics of an organism. RNA editing involves the addition or deletion of cytidine or uridine nucleotide from the mRNA and causes a change in the nature of the codon in the mRNA before its translation. During RNA editing, the addition or deletion of a nucleotide is facilitated with the help of an RNA called guide RNA (gRNA). Often, organellar mRNA undergoes editing. In addition to insertion/deletion editing, RNA may undergo other kinds of modifications such as the conversion of cytidine into uridine or the conversion of adenosine into inosine by specific deaminases. These processes are called conversion editing. When adenosine is converted into inosine, it is translated by ribosome as a guanosine, thus, a CAG codon for glutamine becomes CGG after the conversion of adenosine into inosine, and it codes for arginine instead of glutamine. In addition to mRNA, tRNA, ribosomal (rRNA), and micro RNA (miRNA) may undergo editing. Usually, editing of tRNA leads to reading of a stop codon into leucine.

The process of RNA editing not only makes changes in the nature of protein but also presents an exception to the central dogma, it suggests because the direct transfer of information from DNA to RNA into protein.

RNA editing shows that at least in certain instances, proteins are made from information not present in the DNA sequence. Defective RNA editing has been associated with human cancer and with Lou Gehrig's disease, which is also called amyotrophic lateral sclerosis (ALS).

1.3.4 RNA Silencing and Proteomics

In recent years, an entirely new mechanism for gene control has been found to exist in plants, fungi, and animals. This approach involves the silencing of a particular gene-specific message by causing the degradation of mRNA. RNA silencing controls the expression of the resident gene (s), transgene(s), viral-induced gene(s), and transposons. It was discovered first in the petunia when a gene for anthocyanin was introduced to overexpress the color or pigment formation in the petunia flower. However, in such experiments, the expression of both the resident and the introduced transgene for color synthesis was suppressed, and the plant produced white flowers instead. This phenomenon was called as posttranscriptional gene suppression (PTGS). Later, similar gene suppression was found in the fungus Neurospora. It was determined that the introduction of a gene for orange color in Neurospora resulted in the transformants that were white or albino in color. This phenomenon for silencing a gene, such as the gene for pigment formation in Neurospora, is called quelling. It was found that the albino or white color Neurospora transformants did not produce mRNA specific for the color gene. It was also shown that only a part of the transgene containing only up to 130 nucleotides in length and not the entire gene for color was involved in the quelling of the resident gene. Such a transgene in Neurospora was found to quell or suppress the expression of resident genes even in another nucleus when a heterokaryon was constructed between the transformed and the wild-type strains of Neurospora. Later, Neurospora mutant strains were obtained that were defective in quelling; these mutants were called "quelling deficient" (qde). There are essentially three classes of such mutants in Neurospora. In Neurospora, qde-1 encodes for RNA-dependent RNA polymerase (RdRP), which is required for the synthesis of double-stranded RNA dsRNA such as miRNA or siRNA during gene silencing. The qde-2 gene encodes for the Piwi/Sting class of proteins related to a translation factor eIF2C. The Neurospora qde-3 gene encodes for a protein belonging to the group of WRN (Warner's syndrome) with RNase and DNA helicase functions similar to RecQ DNA helicase. The equivalents of Neurospora qde-1, qde-2, and qde-3 genes have been found to exit in different organisms, including Arabidopsis, worms (*Coenorobdytis elegans*), and fission yeast. Proteins belonging to the RdRP family have

been well characterized from many plants, including tomato, wheat, petunia, and fission yeast, as well as from *C. elegans*. This protein is responsible for making a complementary copy of gene-specific mRNA. This copy of RNA hybridizes with mRNA to form a double-stranded RNA. The latter is degraded to smaller RNA fragments by an enzyme called Dicer, which is similar to RNase III ribonuclease. The RNA fragments then bind to an RNA-induced silencing complex (RISC) and cleave the mRNA specific to a particular gene, which causes gene silencing or suppression.

The role of dsRNA in silencing became obvious from the experiments with worms and Drosophila, and now it is found in mammalian cells as well. It was shown that the introduction of small pieces of dsRNA specific to a gene can cause the degradation of its mRNA, which leads to the suppression of the expression of that gene. Thus, RNA silencing can be used to manipulate the expression of genes in organisms and promise to serve as a great tool in the control of several human diseases, including cancer. Fire and Mello received the Nobel Prize in 2006 for elucidating the mechanism of RNA interference (Fire et al. 1998). It is known that certain infectitous agents, including viruses, trypanosomes, and intestinal parasites, cause havoc in humans because of their ability for antigenic variations. However, it is now known that certain intestinal parasites of humans such as *Giardia lamblia*, maintain their antigenic variation by the use of RNA interference (Prucca et al. 2008). Understanding this process may provide a clue to controlling several infectious diseases in human.

1.4 MOLECULAR BIOLOGY OF GENES AND PROTEINS

A gene is defined as a DNA segment or a stretch of nucleotide sequence that encodes a protein through the process of transcription and translation. However, a few genes make RNA that are not translated into proteins. These genes during transcription make rRNA and tRNA, which facilitate the translation of the transcripts from the protein-encoding genes. In prokaryotic genes, the coding segment of DNA is continuous and their transcripts are translated directly into protein without any modification. Thus, in prokaryotes, the transcript is synonymous to mRNA (i.e., the RNA that carries the information for making of a protein through the process of translation on ribosomes). The existence of mRNA in bacterial cells was demonstrated by Volkin and Astrachan (1957), and later the idea that mRNAs carry the information from DNA to ribosomes for translation into proteins was suggested by Brenner et al. (1961). Simultaneously, Marshall Nirenberg and H. G. Khorana elucidated the genetic codes and the mechanism of information storage and transfer as implied by the Watson-Crick structure of

DNA. Khorana and Nirenberg received the Nobel Prize in 1968 for their contributions. Even as early as the 1960s, the heterogeneous size of transcripts in eukaryotes was known. The eukaryotic transcripts were termed "heterogeneous nuclear RNA" (hnRNA) or premRNA. However, the myth about the heterogeneous nature of eukaryotic transcripts was elucidated by the discovery of the split nature of genes in eukaryotes. In the mid-1970s, it became obvious that some genes in eukaryotes have a split structure in which the coding segments of DNA called exons are interrupted by intervening noncoding DNA segments called introns. This conclusion was based on the results of heteroduplex mapping involving the hybridization of the DNA of a gene with mRNA and the visualization of the heteroduplex structure by electron microscopy. In such experiments, when the DNA of a gene was hybridized with the mRNA, certain DNA sequences appeared as loops (Sharp 2005). The appearance of these loops indicated the presence of the intervening noncoding sequences (introns) that were absent from the mRNA. Based on the results of these hybridization experiments, it was concluded that a transcript undergoes splicing events, which lead to an excision of the introns. Thus, the exons are made continuous and only then the message becomes translatable. These observations established a distinction between the structure of a transcript and its mRNA in the eukaryotes. Later, the presence of exons and introns in a gene was confirmed by comparing the DNA sequence of a gene and its mRNA for the chicken ovulbumin gene. With the completion of the genome projects of many organisms, the presence of exons and introns in a gene is readily established by identifying the occurrence of the conserved nucleotides at the exon–intron junctions.

Initially, it was thought that introns are simply excised out from a transcript by splicing, which makes the exons continuous in the mRNA. Later, it was shown that a particular transcript may yield many different mRNAs, which was facilitated by two different mechanisms called alternate and transsplicing. In alternate splicing, the exons are brought together in different combinations. For example, if there are three exons in a gene, an mRNA may contain exon 1 and exon 2, whereas another mRNA from the same gene contains exon 1 and exon 3 so that the two mRNAs will produce entirely different proteins with different amino acids in the C-terminal ends. These two proteins would have entirely different functions contolling different biochemical reactions in the physiology of an organism. Thus, depending on the number of exons, this method of alternate splicing may produce an array of mRNA for entirely different proteins.

It is suggested that in Drosophila, the DSCAM gene may produce more than 38,000 mRNAs encoding different proteins. Among an array of mRNA, not all mRNAs are translatable for a variety of reasons, including the presence of an early stop codon. Alternate splicing may be tissue specific and

may produce proteins with a specific function. It is known that the Bcl-x gene makes a protein that controls programmed cell death or apoptosis. However, this gene makes two different mRNAs via an alternate splicing mechanism. A smaller version of mRNA produces a smaller protein Bcl-x(s), which promotes apoptosis and controls cancer, whereas a larger version of mRNA makes a larger protein that suppresses apoptosis and supports the growth of cancer.

Alternate splicing may involve exon skipping or intron retention. Exon skipping is commonly found in higher eukaryotes. During exon skipping, a particular exon is skipped during splicing. There are several examples of exon skipping, which is used to produce different versions of tropomyosin specific for skeletal muscle, smooth muscle, and brain cells. Exon skipping is found in Drosophila for the control of sex development. Drosophila has a sex lethal gene called sxl; when exon 2 is skipped during splicing, a female-specific sxl protein is produced, which binds with all subsequent transcript of the same gene, causes exision of exon 2 from all mRNAs, and leads to the development of female flies. However, if the male-specific exon 2 is retained in the first round of splicing, it leads to the production of the male-specific sxl protein and causes the development of male flies.

Intron retention results in the production of mRNAs and their encoded proteins of different lengths. Intron retention is commonly found in plants and lower multicellular organisms.

The mechanism of alternate splicing always involves one transcript. As opposed to alternate splicing, transsplicing involves the splicing of exons of two transcripts produced by the same or distinct genes (Figure 1.6). Transsplicing is commonly found in worms such as *C. elegans*. There is some indication that transsplicing may occur in human brain cells.

On average, a protein-encoding gene in humans is roughly 28,000 nucleotides in length, contains approximately 8 exons of 120 nucleotides or more, and contains approximately 7 introns varying in size from 100 to 100,000 nucleotides. Introns are usually several times of exons in length. A human gene on an average produces 3 mRNAs via alternate splicing.

Splicing is facilitated by splicosomes that consist of more than 100 proteins and five small nuclear (sn) RNAs (snRNAs). Certain regulatory proteins called "splicing regulator (SR) proteins" bind to a particular nucleotide sequence in the exon called the exon splicing enhancer (ESE) and recruit splicosomes. The exon may contain an exon splicing suppressor (ESS) sequence, which prevents the splicosome from splicing.

Defective splicing may cause diseases in humans. More than 15% of mutations that cause diseases in humans result in defective splicing. Defective splicing may result in mutations that alter the splice site or the components of splicesosomes, or it may change factors that control splicing.

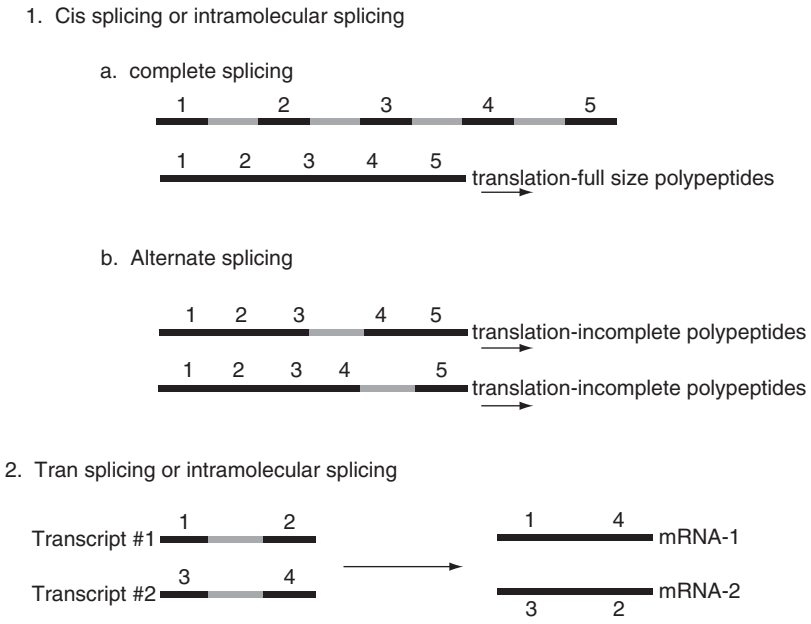

Figure 1.6: Different kinds of splicing of transcripts. (Reproduced from Mishra, 2002 with permission of John Wiley & Sons.)

Many human diseases including cancer may involve mutations that cause defective splicing (Faustino and Cooper 2003). Some genes in which a mutation is known to cause defective splicing and human diseases include BRCA1; BRCA2, HGH, cystic fibrosis; spinal muscular atrophy (SMA), myotonic dystrophy (MD), Wilms tumor suppressor associated with Frasier syndrome (WT1), and many more.

Alternate splicing is the major source of the abundance of proteins in higher organisms. Alternate splicing not only increases the number of proteins but also alters the nature of the protein by insertion and removal of codons in the resulting mRNA. It may also change the reading frame of the mRNA. It could cause termination of protein synthesis by introducing a termination codon in the mRNA. Alternate splicing may control gene expression by changes in the regulatory elements that affect mRNA stability and the translation process.

Alternate splicing has a great effect on speciation as is revealed from the understanding of the genome sequences of humans and mice.

Both have the same number of genes and even share the same exons and introns in many genes. However, it is believed that approximately 25% of the exons that undergo alternate splicing are specific to humans and are different from mice. Likewise, primates have primate-specific alternately

spliced exons that set in the evolution of primates. It seems that the primate-specific exons are derived from mobile genetic elements containing alu sequences. Thus, alu sequences are characteristics of primates. Some of these aspects of the DNA sequence in the human genetic makeup are discussed in Chapter 2.

1.5 PROTEIN CHEMISTRY BEFORE PROTEOMICS

Proteins are described as natural robots, as they seem to know exactly what they have to do within a cell or outside a cell (Tanford and Reynolds 2004). Of course like many other molecules, the function of a protein is determined by its structure. As mentioned, proteins may function in many ways (see Table 1.1). Much of the basic biochemistry of protein was established before the coming of the science of proteomics (see Stryer 1982 and Bell and Bell 1988). This was made possible by developing methods to separate and purify proteins, as well as to determine their specific activity, amino acid composition and sequence, and 3D dimensional structure. Methods were also developed to characterize other physical and biochemical properties, including their regulation and artificial synthesis.

1.5.1 Separation and Purification of Proteins

Proteins were separated from each other during preparation of the cellular extract. Several methods are available to extract proteins from cells or tissue. Proteins are separated by the precipitation in different concentrations of ammonium salts usually in a stepwise manner. Partially purified proteins are separated based on differences in their molecular weights and charges. A small amount of proteins is usually purified based on differences in the molecular weights by ultracentrifugation in a sucrose gradient. Alternatively, they are separated by the method of gel filtration, which acts as a molecular sieve to separate protein molecules based on their sizes. Sepharose (Sephadex; Sigma-Aldrich, St. Louis, MO) is commonly used as a molecular sieve to separate protein molecules of different sizes. Proteins are also separated based on their net positive or negative charges by ion-exchange chromatography. Celluloses such as carboxymethyl (CM) cellulose and diethylaminoethyl (DEAE) cellulose are used in such an ion-exchanger matrix. Several other chromatography methods have been developed that separate protein molecules by their sizes as well as by their charges. Besides chromatography on a solid matrix such as sepharose methods for liquid and high-performance liquid chromatography (HPLC) also

have been developed. A large number of proteins has been purified to homogeneity. Several proteins have been crystallized, and their three-dimensional (3D) structures have been determined.

In addition to different methods of chromatography, several methods of electrophoresis have been developed to separate protein molecules both based on their mass as well as on their charges on a regular gel or on a capillary gel by applying an electrical field. Sodium dodecyl sulfate (SDS) is added to the gel matrix to separate proteins of different molecular weights. SDS is highly negatively charged, and in its presence, all proteins in a mixture become equally negatively charged. Thus, during electrophoresis in the presence of SDS, all proteins move in an electrical field based on their molecular sizes and not on their charges. Smaller proteins move much faster than larger proteins during electrophoresis in a gel that contains SDS. A mixture of proteins also can be separated based on net charges by electrophoresis in a gel that contains a mixture of ampholine of different isoelectric points (pIs). These two methods of eletrophoreses in SDS gel and ampholine gel are combined so that a protein mixture is first run in SDS gel and then in ampholine gel to separate them based on molecular sizes and electrical charges. This method is called 2D gel because it separates proteins based on sizes and charges when run in two planes at right angles to each other. During electrophoresis, proteins are obtained as separate spots visualized by coloring with a dye. 2D gel was first used to separate more than 1100 proteins of *E. coli* simultaneously on one gel. The ability of 2D gel to separate the entire protein content of an organism and to provide information about them in one attempt ushered in an era that marked the beginning of the science of proteomics.

1.5.1.1 Specific Activity of Proteins. The specific activity of a protein is defined as the activity of a protein preparation per milligram of that protein. The activity of protein is usually determined as the enzymatic activity, as the ability to bind to a ligand, or as its biological activity. The specific activity of a protein increases with the increase in the purification of a protein. There are several methods to determine the amount of protein in a preparation. The simplest way is to measure the absorption at 280 nm of light. In addition, several colorimetric methods are available, of which the Lowry method and the Bradford method are commonly used (see Bell and Bell 1988).

1.5.1.2 Molecular Weight Determination. The molecular weight of a protein is an important criterion. It provides the idea about the relative size of the protein molecules. The molecular weight is traditionally determined by ultracentifugation or by chromatography through a matrix

such as Sepharose or by the mobility of protein molecules in SDS gel on electrophoresis with reference to known protein markers.

1.5.1.3 Amino Acid Composition. Proteins comprise 20 different kinds of amino acids. It is important to know the relative abundance of these component amino acids in a protein. Knowledge of the numbers of different amino acids is also important in determining the sequence of the amino acids in a protein molecule. To determine the amino acid composition of a protein, it is hydrolyzed in 6 M HCl for a few hours and then separated by electrophoresis or by chromatography. The individual amino acid spots on an electrophoretogram are dyed with ninhydrin to facilitate its visualization. The number of amino acids in each spot is determined by colorimetric because the intensity of dye in each spot is related to the number of amino acids. Alternatively, the amino acids separated by chromatography as eluents are dyed with fluorescent dye, and the number of amino acid in a particular eluant is again determined spectroscopically as the number of amino acid is proportional to the amount of dye absorbed by the amino acids. The whole process is automated, and the commercially available machine called an amino acid analyzer determines the amino acid composition of a protein within a few hours. The amino acid analyzer was first developed at Rockefeller University in New York City.

1.5.2 Amino Acid Sequence

The sequence of an amino acid in a protein is determined sequentially from the N-terminus. The N-terminus amino acid is identified by the Edman degradation reaction developed at Rockefeller University and later automated in Melbourne, Australia, by Edman and his collaborators (1950, 1967).

To sequence a protein, it is usually fragmented into peptides of approximately 50 amino acids by cyanogens bromide cleavage or by tryptic digestion. Peptides are first separated from one another. A particular peptide is then adsorbed on to a solid surface such as glass fiber coated with cationic polymer polybrene. An Edman reagent phenylisothiocyanate (PTH) is added to the adsorbed peptide in a basic buffer solution of trimethylamine. In this solution, PTH reacts with an amino group of the N-terminal amino acid, which is then selectively separated from the peptide by the addition of an anhydrous acid. The modified N-terminal amino acid isomerizes into phenylthiohydantoin. This is washed off and then identified after chromatography. The cycle is then repeated to determine the next N-terminal amino acid in the remaining peptide that is adsorbed on to glass fiber coated with polybrene. The method of Edman degradation is elegant but riddled with

certain limitations. The method will not work if the N-terminal amino acid is blocked or buried in the bulk of protein.

1.5.3 Chemical Synthesis of Protein

Chemical synthesis of protein has a long history. Synthesis of the first dipeptide glycylglycine was accomplished by Emil Fischer in 1901. Later, he synthesized octadecapeptide with a different amino acid sequence consisting of 15 glycine and 3 leucine amino acid residues. During such synthesis of peptide, he could not control the sequence of amino acids. An important advancement in this direction was made by Bergmann and Zervas (1932) in Germany by introducing the methods for protecting the amino group. In 1935, both Bergmann and Zervas joined Rockefeller University and trained several protein biochemists including William Stein and Standford Moore, who were awarded the Nobel Prize in 1972 as mentioned earlier in this chapter. Using this strategy of Bergmann and Zervas, an octapeptide hormone oxytoxin was synthesized in 1954 by du Vigneaud et al. Vincent du Vigneaud won the Nobel Prize in Chemistry in 1955 for the synthesis of oxytoxin. However, these methods for chemical synthesis in the solution phase were time consuming. A major stride in the chemical synthesis of protein was made by Merrifield in 1963 at Rockefeller University by developing solid-phase synthesis. In this method, an amino acid is attached to an insoluble support through its carboxyl end and then is reacted by another amino acid with an activated carboxyl group but is protected by an alpha amino group. The amino group of the dipeptide is then deprotected by the removal of the protecting group at the amino terminal and then reacted by a third amino acid with a protected amino group and a activated carboxyl group, which leads to the synthesis of the tripeptide. This process of protection, activation, and deprotection is continued in a cyclic manner until the synthesis of the entire peptide or protein is completed. During such chemical synthesis, it is important to protect certain reactive side chains of the amino acid. At the completion of the chemical synthesis, all protected groups are deprotected and then the peptide is cleaved off the solid support. Such peptides or proteins are then examined for the biochemical and biological properties to demonstrate their identity with the naturally synthesized protein. Merrifield used this method to synthesize the first enzyme ribonuclease A (RNaseA). The method for the chemical synthesis is now automated completely. The entire peptide synthesis is carried out by a machine developed at Rockefeller University.

It is important to note that the knowledge of the sequence of amino acids in RNaseA was crucial in the chemical synthesis of this enzyme by the Merrifield group. It is also important to note that the biosynthesis

of a protein inside a cell always occurs from the N-terminal amino acid, whereas in the chemical synthesis, the peptide chain grows from the C-terminal amino acid, which is first attached to an insoluble solid support. It is important to note a long-chain protein like RNaseA is first synthesized in vitro as several component peptides and then they are ligated to yield a full-length protein. Usually, an acid-sensitive tert-butoxycarbonyl (Boc) group or a base-sensitive 9-fluorenylmethyloxycarbonyl (Fmoc) group is used to protect the alpha-amino group of the amino acid to be added to the growing chain of peptides during the chemical synthesis. A detailed method of the chemical synthesis of protein is discussed elsewhere (Nilsson et al. 2005).

1.5.4 Protein Engineering

Our ability to synthesize proteins in vivo and in vitro has led to the development of protein engineering technology. Using this technology, proteins of interest with certain desirable properties are produced in abundance. The process of protein engineering uses two different methods, which are not mutually exclusive. In reality, most laboratories use both methods for protein production. The first method is called "rational design." This requires a complete knowledge of the protein structure, which was difficult in the preproteomic era but has become readily available after the development of proteomics. This method uses site-directed mutagenesis and is a cost-effective method. The second method is called "directed evolution". This method mimics the process of natural evolution because proteins of different kinds are produced by random mutgenesis and then screened to select one with desired features. At times, DNA encoding different proteins are spliced to construct an end-product that combines the desirable features of different proteins. The major drawbacks of this strategy are twofold, one that it is a laborious method that involves several constructs and second that it requires high throughput, which is not possible for certain proteins.

1.5.5 Crystal Structure

The analysis of the structure of the protein crystal provides insight into the 3D structure of the protein with respect to the location of every atom one after another in the amino acid string of the protein. The 3D structure of the protein is usually revealed by the X-ray defraction pattern of the protein crystal. The X-ray defraction pattern is generated by scattering the X ray by the electrons in the atom when a beam of X ray is shined on the protein crystal. The X-ray pattern is then subjected to analysis by Fourier transformation to generate the 3D structure. An analysis of the 3D structure protein

was conducted by John Kendrew and by Max Perutz in the Cavendish laboratory some time in the early 1940s. It took them almost 22 years to determine the 3D structure of a small protein myoglobin (Kendrew 1961) and that of hemoglobin (Perutz et al. 1960), for which both Kendrew and Perutz received the Nobel Prize in 1962. After their work, the 3D structure analysis of protein progressed slowly. By 1990, the structure of less than 100 proteins was revealed by determining the X-ray defraction pattern of the protein crystals.

The whole process was sped up by advancement of a new technique called mad, in which a synchatron was used to beam X ray on the protein crystals. This technique readily provided data about the phase of defraction. To generate the 3D structure, both information regarding the amplitude and phase are required.

An earlier phase was determined by the X-ray defraction of protein crystals containing heavy metals at different positions that required the comparison of several X-ray defraction patterns. The process of the 3D structure analysis was sped up by the advances in computing power as well.

In addition to X ray, nuclear magnetic resonance (NMR) is used to determine the 3D structure of small proteins in solution. NMR is good for proteins that cannot be crystallized. NMR also yields the 3D structure of proteins in dynamic state because protein molecules are in solution unlike the 3D structure of a protein crystal.

1.5.6 Active Site and Regulation of Proteins

One important aspect of proteomics is to understand the function of a protein. This is particularly crucial for understanding the role of protein in causing diseases and in developing drugs. Proteins act in several ways in a cell. Most proteins act by catalyzing a biochemical reaction or by binding with certain molecules including other protein molecules. A protein usually contains an active site as a part of its structure; with the development of proteomics, the active site of an enzyme can be determined by bioinformatics using computing software. The active site binds with a substrate during enzymatic reaction and then catalizes the reaction. Two models are used to explain the binding and catalysis of a substrate. The first model is called the "lock and key model," and the second model is called the "induced fit" model. In the first model, the active site and the substrate have a lock and key relationship, which accounts for their specifity. In the induced fit model, the active site is not a rigid structure and suggests certain flexibility in the active site induced by the binding of a substrate. Molecules that mimic the structure of the substrate can bind with the active site and can

block binding with the substrate; this is the basis for enzyme inhibition by certain drugs and the regulation of enzymes, as discussed below.

Such sites may bind with other proteins or with certain other molecules in the case of nonenzyme proteins. Certain drugs that bear a similarity to the structure of the substrate may bind with the active site of protein and may inhibit the enzymatic activity of the protein by obstructing its interaction with the natural substrate. Such inhibitors are known as competitive inhibitors, because they compete with the substrate for the active site of the protein. Unlike the molecules that bind with the active site of a protein, certain other molecules bind with the protein at a site other than the substrate binding site and bring a conformational change to the structure of the protein such that it cannot bind with the substrate anymore. Such inhibitors are known as allosteric inhibitors and as noncompetitive inhibitors, because they do not compete with the substrate for binding with the active site. The competitive and noncompetitive inhibitors are easily distinguished by Michaelis-Menton kinetics. Such kinetic analysis is carried out by plotting 1/V against 1/S, where S represents the substrate concentration and V represents the velocity of the biochemical reaction (see Bell and Bell 1988), which generates a straight line. The allosteric regulation of protein was established by Jacob and Monod (1964) in Paris. With the creation of appropriate mutants of *E. coli*, Jacob and Monod established the role of an allosteric protein or a repressor protein involved in the control of transcription of the Lac operon in this bacterium. Both Jacob and Monod received the Nobel Prize for this work in 1965.

1.5.7 Signal sequence and Protein Targetting

Proteins are synthesized on ribosomes and then move from their place of syntheses to the different parts of the cell to take residence there and to function in different ways. In 1970s, Gunter Blobel of Rockefeller University identified about 15 amino acids long sequences in different proteins that targets these proteins to their destinations into the cell wall, cell membrane and different organelles including nucleus, nucleolus, golgi bodies, mitochondria, chloroplasts and periosomes or to the exterior of the cell. These Signal sequences are like postal codes that help deliver letters to final destinations. The signal sequences are usually located on the N-terminus of the protein. The signal sequences are usually cleaved by a protease after the transport of the proteins. Proteins transported to different locations usually carry signal sequences of different amino acid composition for example proteins that are destined to endoplasmic reticulum consist of signal sequences of 5-10 hydrophobic amino acids at the N-terminus whereas the signal sequences of those proteins being transported to nucleus contain plus-charged amino

acids within the peptide. The mitochondrial targeting signals contain alternating sequence of hydrophobic and plus-charged amino acids. The proteins being targeted to peroxisomes usually carry a signal sequence of three amino acids on the C-terminus. These proteins destined for transport are usually unfolded and are escorted by a chaperon protein. After the transport is complete the unfolded proteins are folded to assume tertiary structures with the help of a chaperon protein. At times protein can find its destination upon glycosylation i.e. acquisition of carbohydrate moiety.

A genetic defect in this protein transport leads to a number of human diseases. Therefore the understanding of protein transport has been instrumental in understanding these human diseases and may provide clue for their therapy. Blobel was awarded Nobel Prize in 1999 for his work elucidating the mechanism of protein transport in the cell.

The views regarding the general distribution of proteins and enzymes inside the cell, have changed over the years. Earlier it was thought that enzymes are randomly distributed in the cytosol and the enzymatic reactions happened by the chance meeting of an enzyme with the substrate molecule. Contrary to this view, the enzymes of same or related metabolic pathways are localized together and not randomly distributed in the cytosol. They occur together in the vicinity of each other by virtue of certain structural similarities which help them in recognizing each other. This view is also supported by the study of protein-portein interactions discussed in Chapter 5.

1.5.8 Intein

Inteins are segments in a protein that are self-exised out, followed by the joining of remaining segments called the exteins. After the removal of intein, the N-terminal and C-terminal exteins are joined by a peptide linkage as soon as the peptide is synthesized from the mRNA. Inteins in proteins are like introns in genes; intein must be removed to provide a functional protein just as an intron must be removed from a transcript to give a translatable message. Currently, more than 200 inteins have been described from different proteins; a data bank of inteins is available. Inteins are usually 100–800 amino acids in length. Some inteins may be derived from two genes encoding them; for example, the dnae DNA polymerase (dnaE) of cyanobacteria contains two segments, an N-intein segment of 123 amino acids and a C-intein sement of 36 amino acids. The two segments are encoded by two separate genes dnaE-n and dnaE-c for the alpha subunit of the DNA polymerase III. This is equivalent to transsplicing in the case of the genes. The gene-encoding inteins usually carry an endonuclease that helps in the propogation of inteins. Inteins have been found in all forms of

life, including archea, bacteria, and eukaryotes. Inteins have been used in different ways, such as protein engineering and marking a protein for NMR characterization. Inteins may provide a useful tool to develop a drug that can stop the removal of intein from a protein, which renders that protein nonfunctional and, therefore, responsible for the cause of a disease.

1.5.9 Unstructured Protein

Now it is established that there are two classes of proteins in the living systems: one class with an ordered structure and a second class without any ordered structure, intrinsically unstructured, or unordered (Dyson and Wright 2005).

The structure of the first group of proteins was well established before the development of proteomics. All proteins are known to assume several levels of organization, such as the primary, secondary, and tertiary structures. Several proteins have another level of organization called the quaternary structure. The sequence of amino acids in a protein represents the primary structure. The secondary structure represents the coiled structure of the protein because of the folding in the primary structure based on interactions of the amino acids, particularly their side chain among themselves. The tertiary structure of a protein is its 3D structure based on the complete folding of the polypeptide on itself. The tertiary structure determines the final shape of the protein and its activity. Many proteins, after assuming a tertiary structure, interact either with themselves or with another protein to assume a quaternary structure. Proteins with identical peptides as subunits in the quaternary structure are called homomers, whereas those with different peptides as a subunit in the quaternary structure are called heteromers. Hemoglobin that contains two alpha chains and two beta chains is an excellent example of a protein with a quaternary structure. No proteins exist in its primary structure as a straight stretch of peptide as soon as the proteins are synthesized because of the immediate biochemical interactions among the different amino acids in the stretch of polypeptide. The secondary structure is usually determined by circular diachroism or even by gel filtration. The 3D structure of a protein with its tertiary or quaternary structure is best determined by X-ray crystallography or by NMR spectrometry as discussed in Chapter 3. The different structures of a protein can be determined several ways. Not all proteins assume a 3D structure. It is estimated that more than 35% of proteins found in living systems have no intrinsic structure, as they lack a tertiary structure. The proteins are called intrinsically unstructured proteins (Dyson and Wright 2005). These proteins usually lack the bulky hydrophobic amino acids in their primary structure. These proteins exist as random coil chain and are short lived. They usually perform

regulatory functions, such as controlling the regulation of cell cycle, regulating transcription and translation, and signaling pathways. The study of these proteins is in infancy, but it is expected to throw much light on how the function of proteins is regulated.

1.5.10 Protein Misfolding and Human Disease

In addition to the random coil structure of the intrinsically unstructured proteins, other proteins also may assume such an unspecified structure because of mutation or other changes in the proteins. Proteins in several neurodegenerative and other diseases are altered in their structure. They lack the usual secondary and tertiary structure of the protein. They usually result from the misfolding of proteins. As mentioned, Anfinsen established that the secondary and tertiary structures are controlled by the primary structure of the protein. He demonstrated that the enzyme ribonuclease unfolds under denaturing conditions, such as during the addition of urea to the solution that contains the protein. Unfolded ribonuclease loses its enzymatic activity. However, as urea is removed from the protein solution by dialysis, ribonuclease starts folding and assumes a completely folded structure with a secondary and tertiary structure and the full restoration of its enzymatic activity. Anfinsen won the Nobel Prize in 1972 for this work. However, many proteins cannot fold, as they are synthesized inside the cell and remain misfolded. The existence of such misfolded proteins causes several diseases in humans. Certain neurogenerative diseases, including Alzheimer disease, Creutzfeldt-Jakob disease, Kuru, and mad cow disease, result from the misfolding of proteins.

In the 1960s, some proteins that cause mad cow disease were characterized as infective protein molecules. Because they possessed proteins exclusively, these were called prions, which are analogous to virions; the nucleic acid contained the infecting particles. Later, Prusiner, who characterized the prions as protease-resistant proteins (PrPs), was awarded the 1997 Nobel Prize in Physiology and Medicine for his work. The gene for the normal PrP after mutation causes prions in which the mutant PrP cannot fold properly. Later, these misfolded PrPs were shown to be infective protein molecules that perpetuate by causing the misfolding of proteins that otherwise would have exited as normal properly folded proteins with full biological activity. Several other diseases such as cystic fibrosis also result from the misfolding of a protein called the cystic fibrosis transmembrane conductance regulator (CFTR). These misfolded proteins remain in the random coil position and lose the biological activity required for the transport of chloride ions. However, in Alzheimer disease, they become sticky and form the characteristic plaques of beta sheets in the brain of the patients.

When a protein molecule gets misfolded for some reason, it is usually degraded. However, sometimes it escapes degradation and then it acts like a cheparon and causes the misfolding of the other protein molecules. This is the basis of the so-called infectivity of prions that causes mad cow disease and Creutzfeldt-Jakob disease. Once the healthy cow is exposed to misfolded protein or prions, it leads to the misfolding of the other proteins in the brain cells, which causes the disease.

When prions were first discovered, they were considered as infectious protein particles, and for a while, it was thought that the prions acted as a proteinaceous infecting agent parallel to the infective viral DNAs/RNAs. This view presented a challenge to the age-old dogma that only nucleic acids acted as genetic material. However, with the understanding that the formation of new prion particles is induced by the misfolding of other naturally occurring proteins, the myth of a protein as genetic material has been resolved. The prions cannot replicate and thus do not code for the daughter prions. Instead, these prions recruit new prion particles by inducing misfolding of the newly synthesized proteins encoded by the host genome. For example, in cystic fibrosis, a misfolded CFTR protein leads to the misfolding of other CFTR proteins, and the cell loses its normal function. It is shown that the normal PrP controls the long-term memory in mammals. Prions or prion-like particles have been found to exit in yeast and fungi-like Podospora, where they control different phenotypes in the organisms that harbor them. Thus, the understanding of proper folding of protein is crucial in knowing the cause of these diseases and their treatments. Now, a yeast heat shock protein Hsp40/YdjI has been identified that suppresses the aggregation of misfolded proteins and helps in refolding misfolded proteins. It seems to recognize certain repeat sequences as a consensus motif in the protein. The *E. coli* protein Dnaj is homologous to yeast protein. This protein may be useful in understanding human diseases that involve the misfolding of proteins.

REFERENCES

Anfinsen C. B. 1973 Principles that govern the folding of protein chains. Science 181, 223.

Avery, O. T., C. M. MacLeod, and M. McCarthy. 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types I induction of transformation by DNA from Pneumococcus type III. J. Exp. Med. 79, 137

Ayala, F. J. and J. A. Kiger. 1984. Modern Genetics, Second Edition. Menlo Park, CA: Benjamin Cummings.

Beadle, G. W. and E. L. Tatum. 1941. Genetic Control of Biochemical Reactions in Neurospora. Proc. Nat. Acad. Sci. U. S. A. 27 (11), 499–506.

Bell, J. E. and E. T. Bell. 1988. Proteins and Enzymes. Englewood Cliffs, NJ: Prentice Hall.

Bergmann, M. and Zervas L. 1932. Über ein allgemeines Verfahren der Peptid-Synthese. *Berichte der Deutschen Chemischen Gesellschaft* 65(7), 1192–1201.

Berzelius, J. J. 1838. The word "protein" was coined from Greek word proteios meaning the first by Jöns Jakob Berzelius in 1838 in a letter to his friend.

Brenner, S., F. Jacob, and M. Meselson. 1961. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. Nature 190, 576–581.

Crick, F. H. C. 1958. Biosynthesis of macromolecules. Symp. Soc. Exp. Biol. XII, 138–163.

Crick, F. H. 1970. Central dogma of molecular biology. Nature 227, 561–563.

Darwin, C. 1859. On the Origin of Species, 1st ed. London, UK: John Murray.

Dyson, H. J. and P. E. Wright. 2005. Elucidation of the protein folding landscape by NMR. Methods Enzymol. 394, 299–321.

du Vigneaud, V., C. Ressler, J. M. Swan, C. W. Roberts, and P. G. Katsoyannis. 1954. Oxytocin: synthesis. *J. Am. Chem. Soc*. 76; 3115–3118.

Edman, P. 1950. Method for determination of the amino acid sequence in peptides. Acta Chemica Scandinavia. 4, 283–284.

Edman, P. and G. Begg. 1967. A protein Sequenator Eur. J. Biochem. 1, 80–91.

Faustino, N. A. and T. A. Cooper. 2003. Pre-mRNA splicing and human disease. Genes Dev 17, 419–437.

Fire, S. Q., M. K. Xu, S. A. Montgomery, S. E. Kostas, and C. C. Driver. 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature. 391, 806–811

Fischer, E. 1902. *Nobel Lectures, Chemistry 1901–1921* , Amsterdam, The Netherlands: Elsevier 1966

Gorrod, A. E. 1909. Inborn Errors of Metabolism. Oxford, UK: Oxford University Press.

Hershey, A. D. and M. Chase. 1952. Independent functions of viral protein and nucleic acids in growth of bacteriophage. J. Gen. Physiol. 36, 39–56

Ingram, V. M. 1956. A specific chemical difference between globins of normal and sickle-cell anúmia hemoglobins. Nature 178, 792–794.

Ingram, V. M. 1957. Gene mutations in human hemoglobin: the chemical difference between normal and sickle húmoglobin. Nature 180, 326–328.

Jacob, F., and J. Monod. 1964. Biochemical and genetic mechanisms of regulation in the bacterial cell. Bull. Soc. Chim. Biol. 46, 1499–1532.

Kendrew, J. 1961. The three-dimensional structure of a protein molecule. Sci. Am. 205, 96–110.

Khorana, H. G. 1968 Nucleic acid synthesis in the study of Genetic code. Nobel Lecture 341–366.

Klose, J. 1975. Protein mapping by combined isoelectric focusing and electrophoresis in mouse tissues. A novel approach to testing for induced point mutations in mammals. Humangenetik 26: 231–243.

Lander, E. et al. 2001. International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Leder, P., and M. W. Nirenberg. 1964. RNA codewords and protein synthesis3. On the nucleotide sequence of a cysteine and leucine RNA code words. Proc. Na. Acad. Sci. U.S.A. 52, 1521–1529.

Lewin, B. 2004. Gene VIII. Upper Saddle River, NJ: Prentice Hall.

Mattick, J. S. 2003. Challanging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. BioEssays 25, 930–939.

Mendel, J. G. 1866. Versuche über Plflanzenhybriden Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr, 1865. Abhandlungen, 3–47. For the English translation, see: Druery, C. T. and W. Bateson. 1901. Experiments in plant hybridization. J. R. Horticul. Soc. 26, 1–32.

Merrifield, R. B. 1963. Solid Phase Peptide Synthesis. I. The Synthesis of a Tetrapeptide. J. Am. Chem Sci J. 85(14), 2149–2154.

Mishra, N. C. 2002. Nucleases—Molecular Biology and Applications. New York: Wiley.

Mitchell, H. K. and J. Lein. 1948. A Neurospora mutant deficient in the enzymatic synthesis of tryptophan. J. Biol. Chem. 175, 481–482.

Mitchell, H. K., M. B. Houlahan, J. Lein. 1948. Some aspects of genetic control of tryptophan metabolism in Neurospora. Genetics 33, 620.

Moore, S. and W. Stein. 1972. The chemical synthesis of pancreatic ribonuclease and deoxyribonuclease. Nobel Lecture 80–93.

Nilsson, B. L., M. B. Soellner, and R. T. Raines. 2005. Chemical synthesis of proteins. Annu. Rev. Biophys. Biomol. Struct. 34, 91–118.

O'Farrell, P. H. 1975. High resolution two-dimensional electrophoresis of proteins. J. Biol. Chem. 250, 4007–4021.

Pauling Linus 1954 in Nobel Lectures, *Chemistry 1942-1962*, Elsevier Publishing Company, Amsterdam, 1964.

Perutz, M. F. Rossman, MG; Cullis, AF; Muirhead, H; Will, G; North, 1960. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5A resolution, obtained by x-ray analysis. Nature 185, 416–422.

Prucca, C. G., I. Salvin, R. Quiroga, E. V. Elias, F. D. Rivero, A. Saura, P. G. Carranza, and H. D. Lujan. 2008 Antigenic variation in Giardia lamblia is regulated by RNA interference. Nature 456, 750–754.

Sanger, F. 1952. The arrangement of amino acids in proteins. Adv. Protein Chem. 7, 1–28

Sanger, F. 1958. The chemistry of insulin. Nobel Lecture 544–556.

Sarabhai, A. S., A. W. O. Stretton, S. Brenner, and A. Bolle. 1964. Colinearity of the gene with the peptide chain. Nature 201, 13–17.

Sharp, P. A. 2005. The discovery of split genes and RNA splicing. Trends Biochem Sci. 30, 279–81.

Stryer, L. 1982. Biochemistry, 2nd edition. San Francisco, CA: W.H. Freeman Co.

Sumner, J. B. 1946. The chemical nature of enzyme. Nobel Lectures 114–121

Tanford, C. and J. Reynolds. 2004. Nature's Robot: A History of Proteins. Oxford, UK: Oxford University Press.

Volkin, E. and L. Astrachan. 1957. Phosphorus incorporation in Escherichia coli ribo-nucleic acid after infection with bacteriophage T2. Virology 1956, 149–161.

Watson, J. D. and F. H. C. Crick 1953a. Molecular structure of nucleic acids: A structure for desoxyribonucleic acids. Nature 171 731

Watson, J. D. and F. H. C. Crick 1953b. General implications of the structure of desoxyribonucleic acids. Nature 171. 964

Watson, J. 1965. Molecular Biology of Gene. Melno Park, CA: W. A. Benjamin.

Wilkins, M. 1996. 1997. Protein identification in the post-genome era: the rapid rise of proteomics. Q. Rev. Biophys. 30(4), 279–331.

Yanofsky, C. 1952. The effect of gene changes on tryptophan desmolase formation. Proc. Nal. Acad. Sci. U.S.A. 38, 215–226.

Yanofsky, C., B. C. Carlton, J. R. Guest, D. R. Helinski, and U. Henning. 1964. On the colinearity of gene structure and protein structure. Proc. Nat. Acad. Sci. U.S.A. 51, 266–27

Yanofsky, C. 2005a. The favorable features of Tryptophan synthetase for proving Beadle and Tatum's one gene—one enzyme hypotheis. Genetics 169, 511–516.

Yanofsky, C. 2005b. Using studies on Tryptophan metabolism to answer basic biological questions. J. Biol. Chem. 278, 19859–10878.

FURTHER READING

Anraky, Y., R. Mizutani, and Y. Satow. 2008. Protein splicing: Its discovery and structural insight into novel chemical mechanisms. IUMBMB 57, 563–574.

Baltimore, D. 1971. RNA viruses. Bacteriol. Rev. 35, 235.

Baltimore, D. 1975. Viruses, Polymerase and Cancer. Nobel Lectures. Amsterdam, The Netherlands: Elsevier, pp. 215–226.

Beadle, G. W. 1958. Genes and chemical reactions in Neurospora. Nobel Lecture. 587–599.

Cech, T. R. 1988. Conserved sequences and structure of group I intron: Building an active site for RNA catalysis—a review. Gene 73, 259.

Chou, K. C. and Y. D. Kai 2004. A novel approach to predict active sites of enzyme molecules. Proteins 55, 77–82.

Hozumi, N. and Tonegawa, S. 1976. Evidence for rearrangement of immunoglobulin genes coding for Variable and constant regions Proc. Natl. Acad. Sci. 73, 203–207.

Hozumi, N., and S. Tonegawa. 1976. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. Proc. Natl. Acad. Sci. U.S.A. 73(10), 3628–3632.

Kaiser, A. D., and D. S. Hogness. 1960. The transformation of Escherichia coli with deoxyribonucleic acid isolated from bacteriophage lambda-dg. J. Mol. Biol. 392–415.

Keta, P., D. W. Summers, H.-Y. Ren, D. M. Cyr, and N. W. Dekhelyan. 2009. Identification of a concensus motif in substrates bound by a TypeI Hsp40. Proc. Nat. Acad. Sci. U.S.A. 106, 1–5

Margulies, E. H., et al. 2005. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. Proc. Nat. Acad. Sci. U.S.A. 102, 4795–4800.

Nguyen, H. D. and C. K. Hall. 2004. Molecular dynamics simulation of spontaneous fibril formation by random coil peptide. Proc. Natl. Acad. Sci. U.S.A. 101, 16180–16185.

Stahl, N. and S. B. Prusiner. 1991. Prions and prion proteins. FASEB J. 5, 2799–2807.

Tatum, E. L. 1958A. Case history in biological research. Nobel Lecture 2–9.

Taubes, G. 1996. Misfolding the way to disease. Science 271, 1493–1495.

Thomas, P. J., B-H Qu, and P. L. Peterson. 1995. Defective protein folding as a basis of human disease. TIBS 20, 456–459.

Temin, H. 1975. DNA provirus hypothesis. Nobel Lecture 215–245

Tonegawa, S. 1987. Somatic generation of immune diversity. Nobel Lecture 380–405

Venter, J. C., et al. 2001. The sequence of the human genome. Science 291, 1304–1351.