MODULE VI

Protein and RNA structure Prediction: Predicting RNA secondary structure - Nussinov Algorithm, Energy minimisation methods - Zuker Algorithm.

Amino Acids, Polypeptide Composition, Protein Structures, Algorithm for protein folding, Structure prediction

RNA

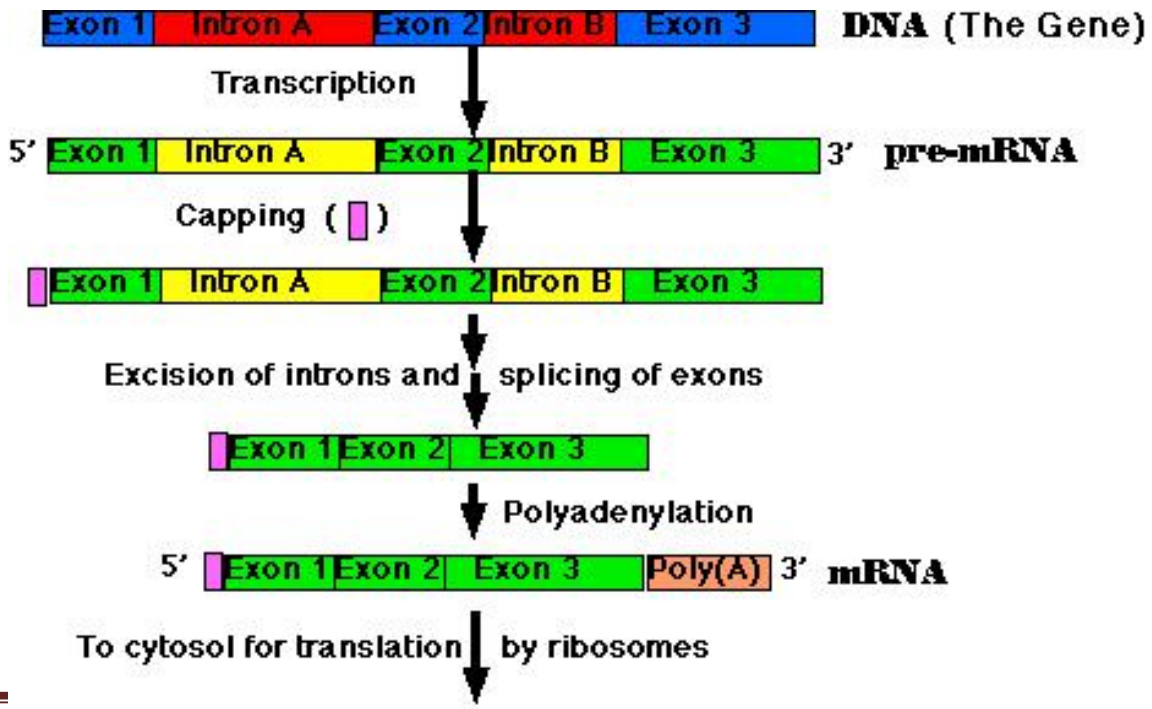
RNA, DNA and proteins are the basic molecules of life on Earth. Recall that:

DNA is used to store and replicate genetic information, proteins are the basic building blocks and active players in the cell, and RNA plays a number of different important roles in the production of proteins from instructions encoded in the DNA.

In eukaryotes, DNA is transcribed into pre-mRNA, from which introns are spliced to produce mature mRNA, which is then translated by ribosomes to produce proteins with the help of tRNAs. A substantial amount of a ribosome consists of RNA.

The RNA-world hypothesis suggests that originally, life was based on RNA and over time RNA dele-gated the data storage problem to DNA and the problem of providing structure and catalytic functionality to proteins.

Below you see the process of transcription.



Structure of RNA

Primary structure

- Primary structure consists of a linear sequence of nucleotides that are linked together by phosphodiester bonds .
- RNA is a single strand of nucleotides (bases) adenine (A), guanine (G), cytosine (C) and uracil (U). The sequence of the bases A, G, C and U is called the primary structure of an RNA.
- In RNA, G and C can form a base pair G≡C by a triple-hydrogen bond, A and U can form a base pair A=U by a double-hydrogen bond, and G and U can form a base pair G−U by a single hydrogen bond. Due to these hydrogen bonds, the primary structure of an RNA can fold back on itself to form its secondary structure.

Secondary Structure

In secondary structure the Watson-Crick and wobble base pairs occurring in the RNA fold.

Suppose that we have the following RNA sequence.

A–G–G–C–C–U–U–C–C–U

Then, this primary structure sequence can fold back on itself to form many possible secondary structures. In Figure 6.2, we show six possible secondary structures of this sequences. In nature, however, there is only one secondary structure to correspond to an RNA sequence.

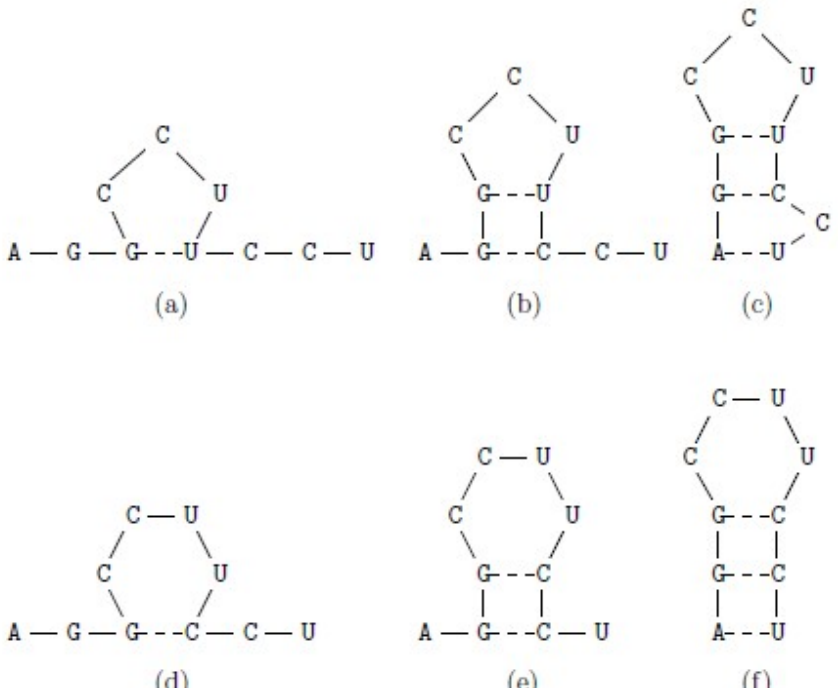


Figure 6.2: Six Possible Secondary Structures of RNA Sequence A–G–G–C–C–U–U–C–C–U

(The Dashed Lines Denote the Hydrogen Bonds)

According to the thermodynamic hypothesis, the actual secondary structure of an RNA sequence is the one with the minimum free energy. In nature, only the stable structure can exist and the stable strucutre must be the one with the minimum free energy. In a secondary structure of an RNA, the base pairs will increase the structural stability, but the unpaired bases will decrease the structural stability. The base pairs of the types G≡C and A=U (called Watson-Crick base pairs) are more stable than that of the type G−U (called wobble base pairs). Together they are all called canonical base pairs.

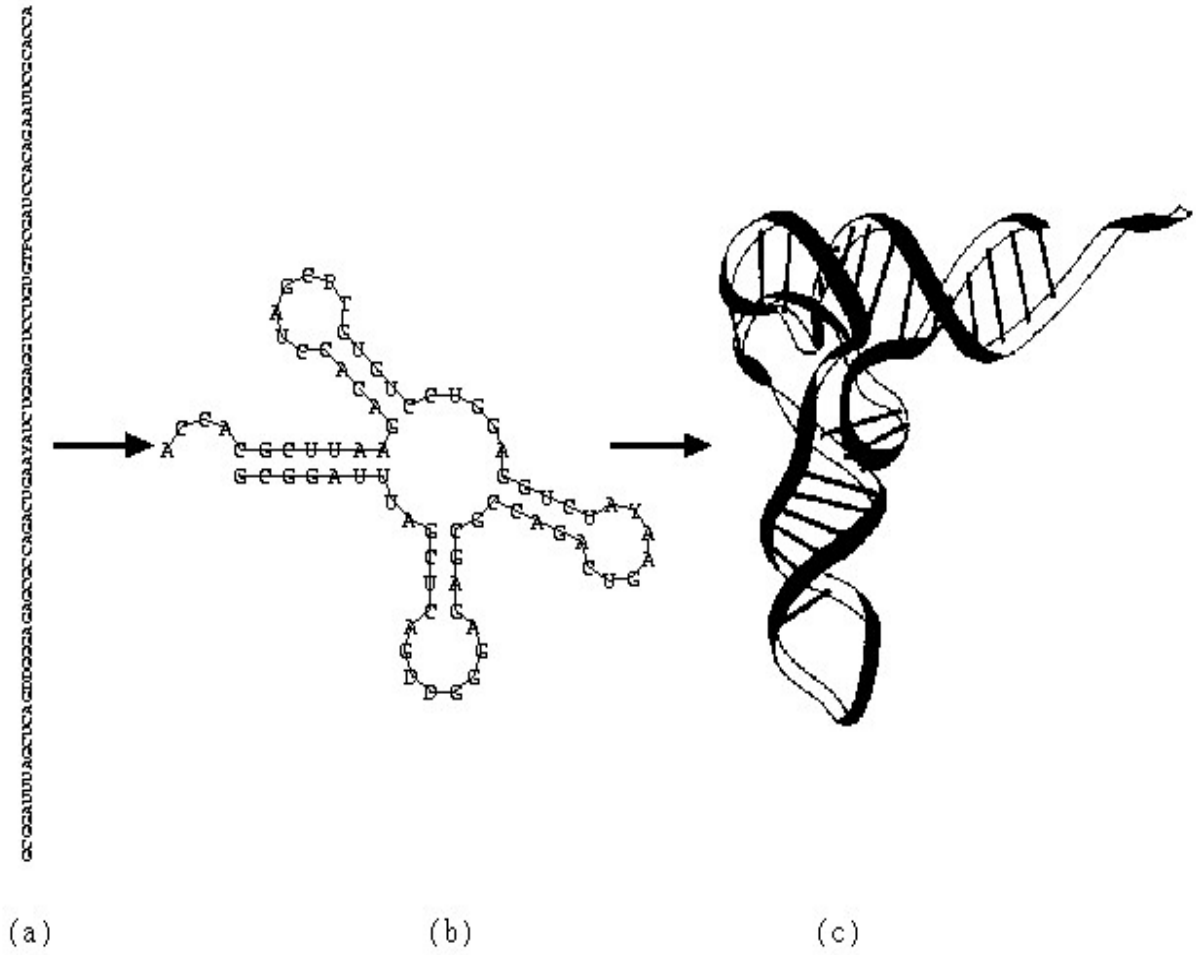
According to these factors, we can find that the sencodary structure of Figure 6.2 (f) is the actual sencondary structure of sequence A–G–G–C–C–U–U–C–C–U.

Hence, we formally define the secondary structure prediction problem as follows. Given an RNA sequence, determine the secondary structure of the minimum free energy from this sequence.

3D structures

There are many different types of RNAs, such as mRNA, tRNA and rRNA each with the different function. It is well known that the function of an RNA is determined by its three-dimensional structure. Hence, knowing the three- dimensional structures of RNAs is important for us to understand their functions.

The three-dimensional structure of an RNA can be predicted experimentally by X-ray crystallography and nuclear magnetic resonance (NMR). But these experimental methods are difficult and quite time consuming. Moreover, they are not always feasible because X-ray crystallography can be applied only to crystallized molecules and NMR is limited to small molecules (< 200 amino acids at present). It was discovered that the three-dimensional of an RNA can be uniquely determined from its sequence (i.e., the primary sequence). Hence, much theoretical effort has been made in determining the three-dimensional structure of an RNA from its sequence alone. Up to now, it is still a hard work to predict the three-dimensional structure of an RNA directly from its sequence. Physical methods like NMR studies to predict RNA secondary structure are expensive and difficult. Computational RNA secondary structure prediction is easier. However, there are efficient algorithms to predict the secondary structure of an RNA, which is useful in predicting the three-dimensional structure. To predict the three-dimesional structure of an RNA sequence, we can first determine its secondary structure and then predict its three-dimensional strucutre according to this secondary structure. RNAStrand folds upon itself to form base pairs & secondary structures



(a) (b) (c)

a) The Primary Structure of the RNA (b) The Secondary Structure of the RNA (c) The Three-Dimensional Structure of the RNA

Hence, we formally define the secondary structure prediction problem as follows. Given an RNA sequence, determine the secondary structure of the minimum free energy from this sequence.[The most common approach to treat RNA structures algorithmically, is to reduce them to the set of base pairs, the so-called *secondary structure*, thereby abstracting from the actual spatial arrangement of nucleotides. For a valid secondary structure, we require that each nucleotide *i* interacts with at most one other nucleotide *j* to form a base pair (i,j). We only consider canonical base pairs, i.e. the Watson–Crick pairs AU, UA, CG, and GC, as well as the so-called wobble pairs GU and UG. Moreover, we usually exclude pseudo-knots (The simplest RNA pseudoknot is formed by base-pairing of nucleotides within a hairpin loop to a complementary sequence outside the hairpin (H-pseudoknot).) i.e. crossing pairs (i,j) and (k,l)where i<k<j<l.

Below figure shows some of the typical configurations in RNA secondary structures.

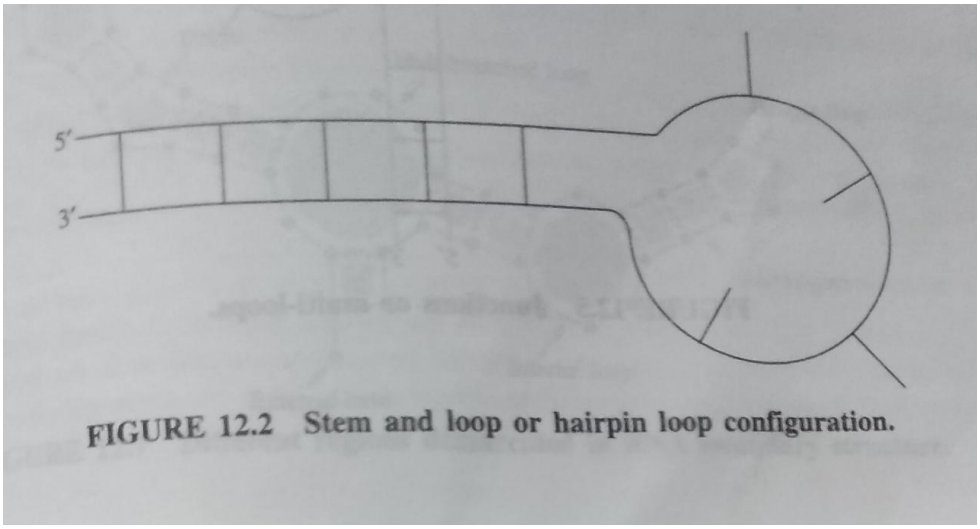
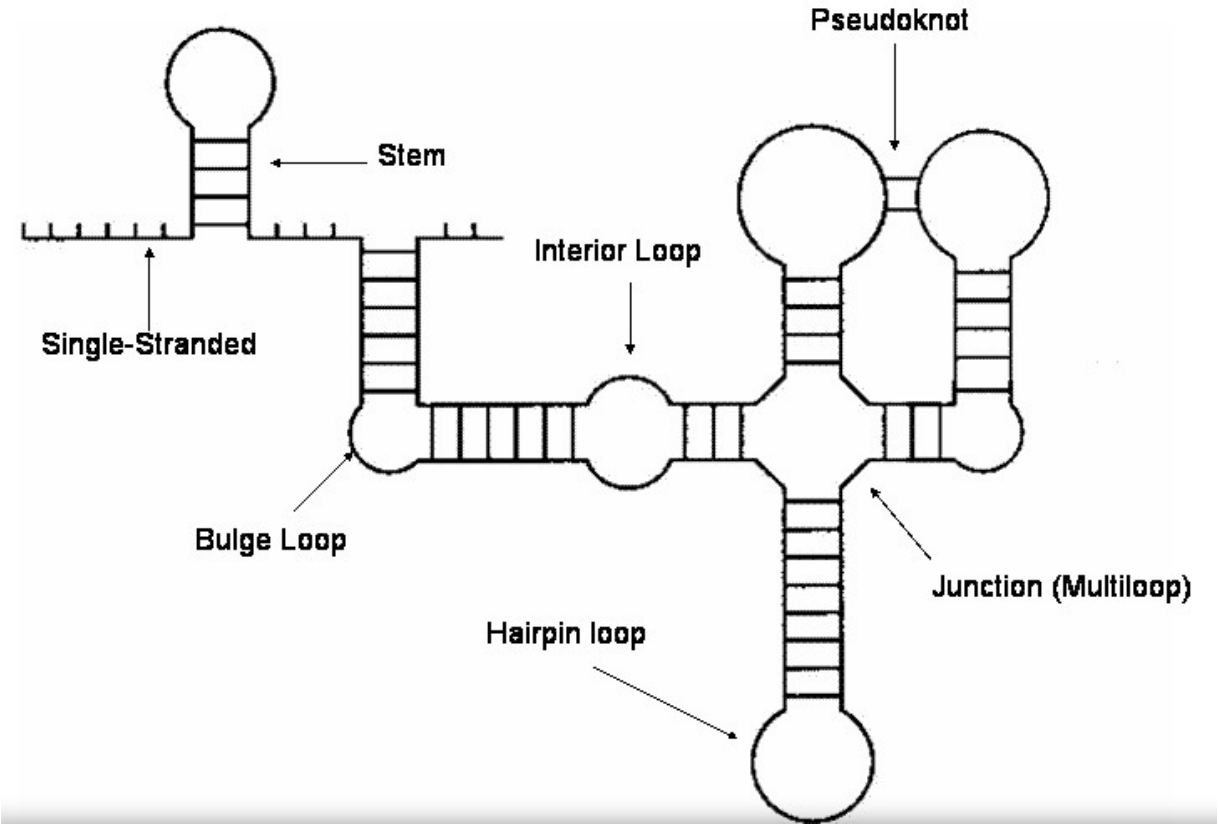


FIGURE 12.2 Stem and loop or hairpin loop configuration.



Secondary Structure of RNA

An RNA sequence will be represented as a string of n characters R = r1r2 · · · rn, where

ri □ {A, C, G, U}. Typically, n can range from 20 to 2000. A secondary structure of R is a set S of base pairs (ri, rj), where $1 \leq i < j \leq n$, such that the following conditions are satisfied.

(1) $j - i > t$, where t is a small positive constant. Typically, t = 3.

(2) If (ri, rj) and (rk, rl) are two base pairs in S and $i \leq k$, then either

(a) i = k and j = l, i.e., (ri, rj) and (rk, rl) are the same base pair,

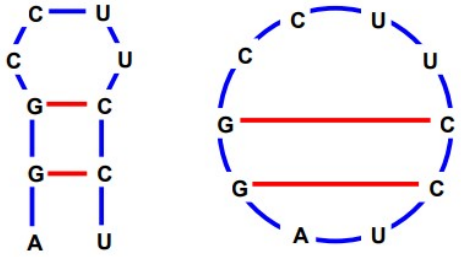
(b) $i < j < k < l$, i.e., (ri, rj) precedes (rk, rl), or

(c) $i < k < l < j$, i.e., (ri, rj) includes (rk, rl).

- The first condition implies that RNA sequence does not fold too sharply on itself.

- The second condition means that each nucleotide can take part in at most one base pair, and guarantees that the secondary structure contains no pseudoknot. Two base pairs (ri, rj) and (rk, rl) are called a pseudoknot if i < k < j < l (see Figure 6.3). Pseudoknots do occur in RNA molecules, but their exclusion simplifies the problem.
- By the above definition, a secondary structure can be represented as an outerplanar graph with degree at most 3, where an outerplanar graph is a graph which can be drawn in the plane in such a way that all vertices (i.e., nucleotides) are arranged on a circle and all edges (i.e.,base pairs) lie inside the circle and do not intersect..

Outerplanar graph: A secondary structure can be represented as an outerplanar graph with degree at most 3. Outerplanar graph: a graph in which all vertices are arranged on a circle and all edges lie inside the circle and do not intersect



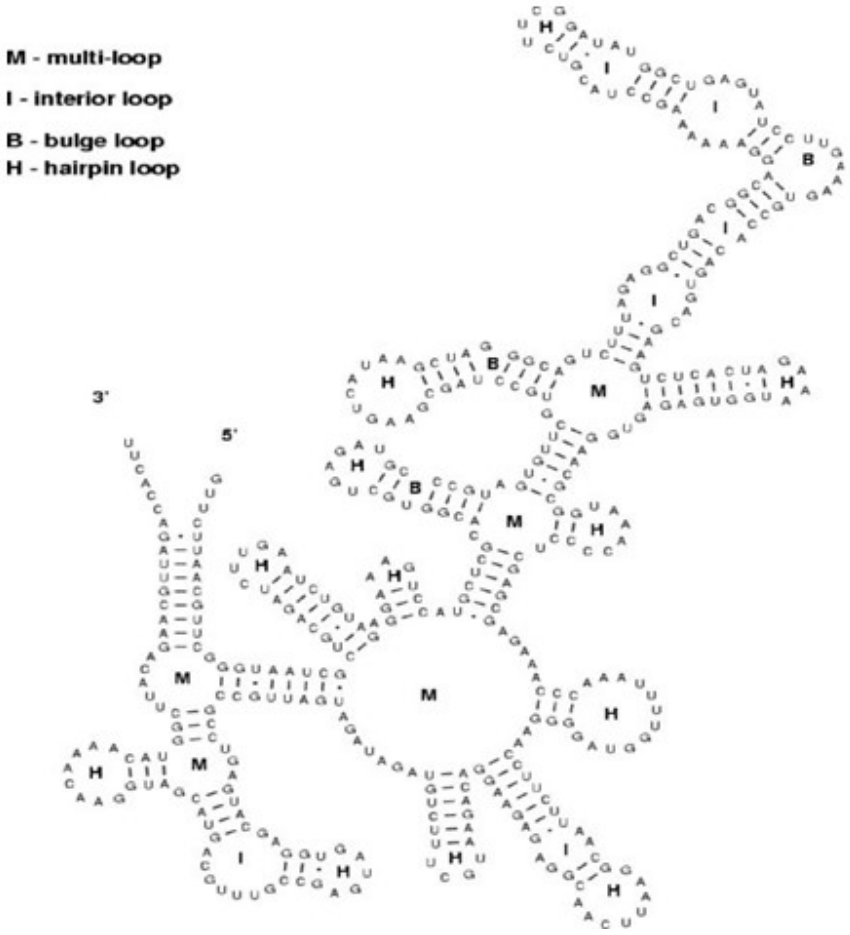
The simplest method of measuring the free energy of S is to assign an energy to each base pair of S and then the free energy of S is the sum of the energies of all base pairs. Due to different hydrogen bonds, the energies of base pairs are usually assigned as different values. For example, the reasonable values for A≡U, G=C and G–U are -3, -2 and -1 (Kcal/mole), respectively. Other possible values might be that the energies of base pairs are all equal. In this case, the problem becomes the one of finding a secondary structure with the maximum number of base pairs. This version of the secondary structure prediction problem is also called RNA maximum base pair matching problem since we can view a secondary structure as a matching.

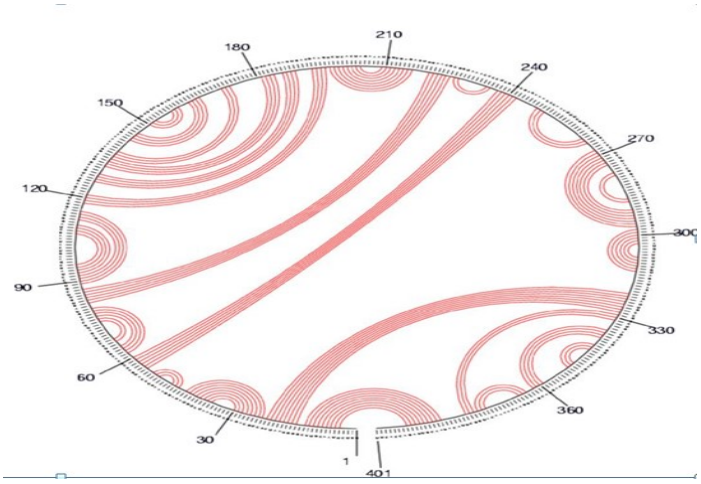
Example of secondary structure

Predicted structure for Bacillus Subtilis RNAase P RNA:

Fig shows two different ways to represent RNA secondary structure

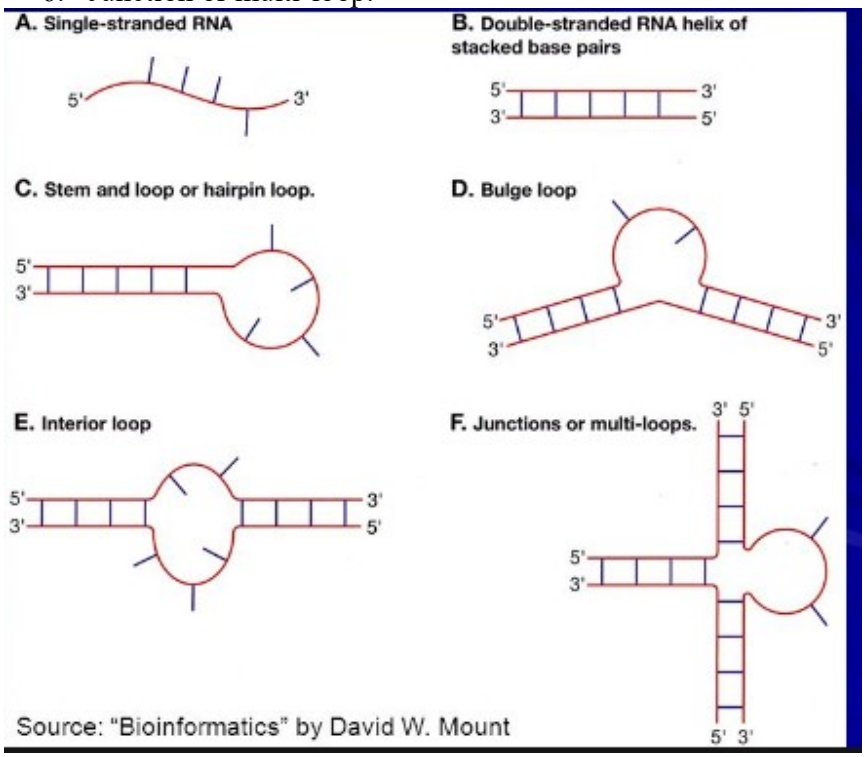
1. Base pair representation
2. Circle representation


M - multi-loop
I - interior loop
B - bulge loop
H - hairpin loop
3'
5'


180
210
240
150
270
120
90
330
60
360
30
1
401

The different types of single and double stranded regions in RNA secondary structure are given in below figure.

1. Single-stranded RNA.
2. Double-stranded RNA helix of stacked base pairs
3. Stem and loop or hairpin loop
4. Bulge loop
5. Interior loop
6. Junction or multi-loop.



RNA structure representation

There are various ways to represent the RNA structure.. The most common ones are:

1. The ”base pair” graph representation
2. The linear representation
3. The mountain representation
4. The bracket representation
5. The circle representation
6. The tree representation

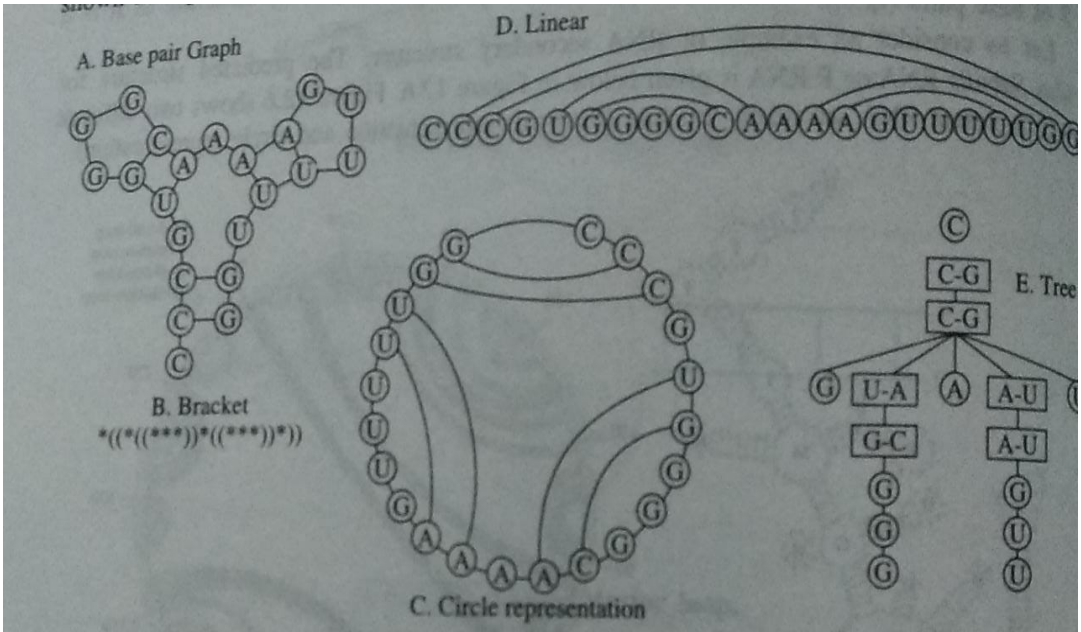


FIGURE 12.8 Different ways to represent RNA structure.

Methods of RNA structure prediction

The problem of predicting the secondary structure of RNA has some similarities to DNA alignment, except that the sequence folds back on itself and aligns complementary bases rather than similar ones.

The goal of aligning two or more biological sequences is to determine whether they are homologous or just similar. In contrast, a secondary structure for an RNA is a simplification of the complex three-dimensional folding of the RNA molecule.

There are two principal approaches for RNA structure prediction. These are

1. Maximize base pair approch

2. Minimize Energy Approch

1. <u>Maximize Base Pairs Approach</u>

This approach is based on a given RNA sequence.

Steps:

1. Determine the set of maximal base-pairs (no base-pair across each other).
2. Then align bases according to their ability to pair with each other.

This gives an approach to determine the optimal structure using **dynamic programming approach or the Nussinov Algorithm**

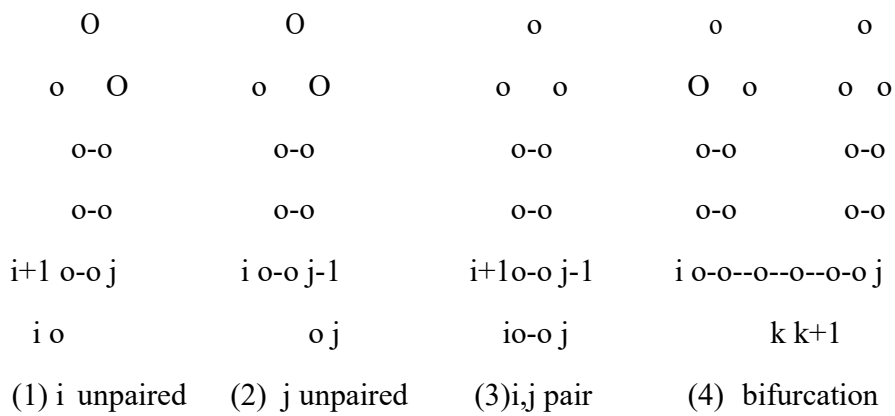
The Nussinov folding algorithm

The simplest approach to predicting the secondary structure of RNA molecules is to find the configuration with the greatest number of paired bases. Note that the number of possible configurations to be inspected grows exponentially with the length of the sequence.

Fortunately, we can employ dynamic programming to obtain an efficient solution. In 1978 Ruth Nussinov et al. published a method to do just that.

The algorithm is recursive. It calculates the best structure for small subsequences, and works its way outward to larger and larger subsequences. The key idea of the recursive calculation is that there are only four possible ways of the getting the best structure for i; j from the best structures of the smaller subsequences.

Idea: There are four ways to obtain an optimal structure for a sequence i; j from smaller substructures:



1. Add an unpaired base i to the best structure for the subsequence [i + 1; j],
2. add an unpaired base j to the best structure for the subsequence [I , j- 1],
3. add paired bases i and j to the best structure for the subsequence [i + 1; j- 1],
4. combine two optimal substructures [i; k] and [k + 1; j].

The dynamic programming algorithm has two stages:

In the fill stage, we will recursively calculate scores (i; j) which are the maximal number of base pairs that can be formed for subsequences (x_i; : : : ; x_j).

In the traceback stage, we traceback through the calculated matrix to obtain one of the maximally base paired structures.

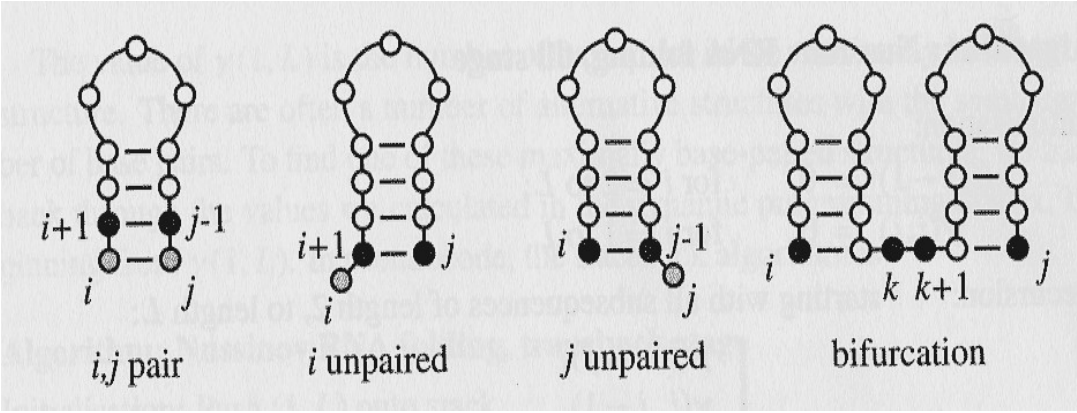
- Find structure with the most base pairs
  - Only consider A-U and G-C and do not distinguish them
- Nussinov algorithm (1970s)

Too simple to be accurate, but stepping-stone for later algorithms

- Problem definition
  - Given sequence $X=x_1x_2\ldots x_L$,compute a structure that has maximum (weighted) number of base pairings
- How can we solve this problem?
  - Remember: RNA folds back to itself!
  - S(i,j) is the maximum score when $x_i..x_j$ folds optimally
  - S(1,L)?
  - S(i,i)?

Grow from Sub structures:

1 2 3 4



$$S(i,j) = max \begin{cases} S(i+1, j-1) + w(i,j) & (1) \\ S(i+1, j) & (2) \\ S(i, j-1) & (3) \\ max_{i<k<j} S(i,k) + S(k+1, j) & (4) \end{cases}$$

$w(i,j) = 1$ if i, j are complementary (i.e., GC, CG, AU or UA); 0 otherwise

- Compute S(i,j) recursively (dynamic programming)
  - Compares a sequence against itself in a dynamic programming matrix

Three steps:

1.Initialization

1. Initialization

	G	G	G	A	A	A	U	C	C
G	0								
G	0	0							
G		0	0						
A			0	0					
A				0	0				
A					0	0			
U						0	0		
C							0	0	
C								0	0

Example:

GGGAAAUCC

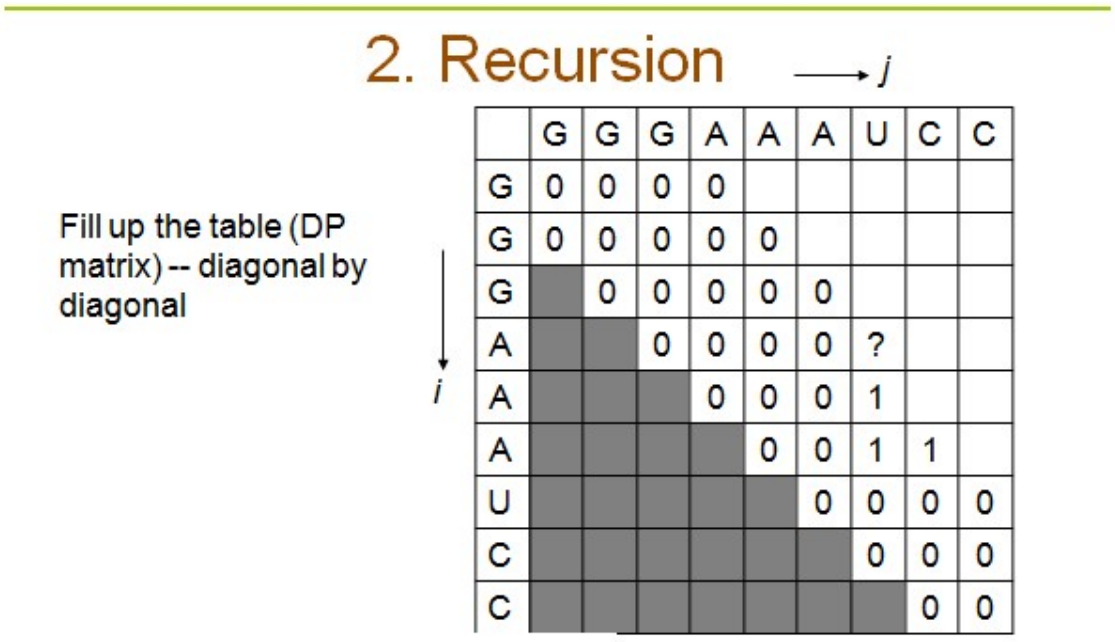
$S(i,i) = 0 \quad \forall \quad 1 \le i \le L$ → the main diagonal

$S(i,i-1) = 0 \quad \forall \quad 2 \le i \le L$ → the diagonal below

L: the length of input sequence

2.Recursion

Fill up the table(DP Matrix) diagonal by diagonl



$$S(i,j) = max \begin{cases} S(i+1, j-1) + w(i,j) & (1) \\ S(i+1, j) & (2) \\ S(i, j-1) & (3) \\ max_{i<k<j} S(i,k) + S(k+1, j) & (4) \end{cases} \quad w(i,j) = \begin{cases} 1 & i, j \text{ are complementary} \\ 0 & otherwise \end{cases}$$

3. Trace back

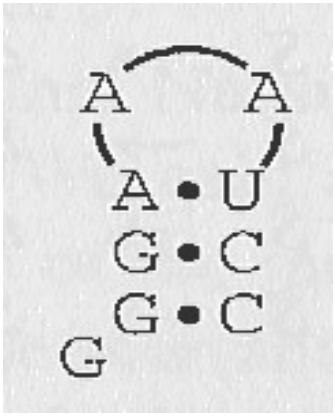
3. Traceback

	G	G	G	A	A	A	U	C	C
G	0	0	0	0	0	0	1	2	3
G	0	0	0	0	0	0	1	2	3
G		0	0	0	0	0	1	2	2
A			0	0	0	0	1	1	1
A				0	0	0	1	1	1
A					0	0	1	1	1
U						0	0	0	0
C							0	0	0
C								0	0

The structure is:

What are the other "optimal" structures?

The structure is:



Maximum basepairs= 3

base-pair #1: 2 (G) - 9 (C)

base-pair #2: 3 (G) - 8 (C)

base-pair #3: 6 (A) - 7 (U)

Minimize Energy Methods

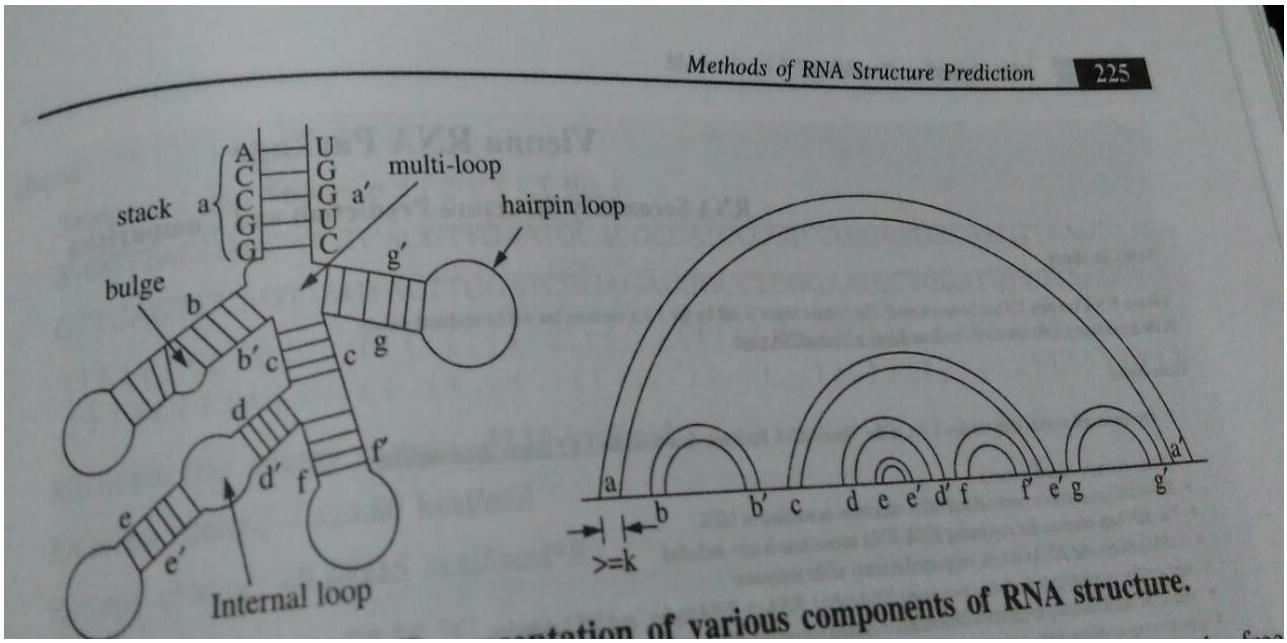
- Nussinov algorithm (base pair maximization) is too simple to be accurate
- In energy minimization, all possible choices of complementary sequences are considered
- Then find the most stable structure.

  Base pairs appears in **'clusters':(STACKS)**

  These are energetically favourable.

  Most of the stability of RNA secondary structure is determined by stacks as they contribute to the negative free energy.

  Unpaired bases form destabilizing loops and these contribute to the positive free energy



of various components of RNA structure.

- Energy minimization algorithm predicts secondary structure by minimizing the free energy (ΔG)
- ΔG calculated as sum of individual contributions of:
  - loops
  - base pairs
  - secondary structure elements

Given the energy tables, the free energy can be calculated for a structure. The score in dynamic programming is based on the free energy values. Gaps represent some form of loop.

RNA secondary structure can be viewed as:

Stacks plus unpaired loops

Energy of the structure=sum of energies of stacks and loops.

Stack configuration can be of different types like:

1. Nested Stacks
2. Parallel Stacks
3. Crossing Stacks(Pseudo notes)

Software packages that incorporates minimum free energy algorithm:

1. **Mfold**(most widely used)
   - MFold predicts optimal and suboptimal secondary structures for an RNA or DNA molecule using the most recent energy minimization method of Zuker.
   - MFold calculates energy matrices that determine all optimal and suboptimal secondary structures for an RNA or DNA molecule. The program writes these energy matrices to an output file. A companion program, PlotFold, reads this output file and displays a representative set of optimal and suboptimal secondary structures for the molecule within any increment of the computed minimum free energy you choose. You can choose any of several different graphic representations for displaying the secondary structures in PlotFold.
2. **Vienna RNA package.**
   - The ViennaRNA Package consists of a C code library and several stand-alone programs for the prediction and comparison of RNA secondary structures.
   - RNA secondary structure prediction through energy minimization is the most used function in the package. We provide three kinds of dynamic programming algorithms for structure prediction: the minimum free energy algorithm of (Zuker & Stiegler 1981) which yields a single optimal structure, the partition function algorithm of (McCaskill 1990) which calculates base pair probabilities in the thermodynamic ensemble, and the suboptimal folding algorithm of (Wuchty et.al 1999) which generates all suboptimal structures within a given energy range of the optimal energy
3. RNAsoft
   - **RNAsoft** is a collection of online services for the computational prediction and design of RNA/DNA structures. The underlying algorithms have been designed and implemented by members of the Bioinformatics, Empirical and Theoretical Algorithmics (BETA) Lab at the Department of Computer Science of the University of British Columbia.
4. BiBiServ

- Is a tool to predict RNA pseudoknots.

Zucker Algorithm

The Zuker algorithm is similar to Nussinov's algorithm: it also employs a dynamic programming approach, using a 4-case recursion equation.

As indicated, the most important difference to the Nussinov calculation is that the Zuker algorithm focuses on loops rather than base pairs. This provides a better fit to experimentally observed data.

We will use two matrices, W and V.

Let $A = (a_1, a_2, \ldots, a_L)$ be a string over the alphabet $\Sigma = \{A, G, C, U\}$.

For $i < j$, let $W(i, j)$ denote the minimum folding energy of all foldings of the subsequence $a_i, \ldots, a_j$.

Additionally, let $V(i, j)$ denote the minimum folding energy of all foldings P of the subsequence $a_i, \ldots, a_j$, with the additional condition that P *contains the base pair* (a_i, a_j).

The following observation is simple, but crucial:

$$W(i, j) \leq V(i, j) \text{ for all } i, j.$$

The two matrices W and V are initialized as follows:

$$W(i, j) = V(i, j) = \infty \text{ for all } i, j \text{ with } j - 4 < i < j.$$

Note that this will to enforce that two paired bases are at least 3 positions away from each other.

10.9.1 Loop-dependent energies

We define different energy functions for the different types of loops:

- Let $eh(i, j)$ be the energy of the hairpin loop closed by the base pair (i, j) (destabilizing, therefore positive),
- let $es(i, j)$ be the energy of the stacked pair (i, j) and $(i + 1, j - 1)$ (stabilizing, therefore negative),
- let $ebi(i, j, i', j')$ be the energy of the bulge or interior loop that is closed by (i, j), with (i', j') accessible from (i, j) (destabilizing, therefore positive), and
- let a denote a constant energy term associated with a multi-loop (destabilizing, therefore positive).

Predicted free-energy value $es(i, j)$ (in kcal/mol at 37^oC) for a stacked base pair (i, j):

	A/U	C/G	G/C	U/A	G/U	U/G
A/U	-0.9	-1.8	-2.3	-1.1	-1.1	-0.8
C/G	-1.7	-2.9	-3.4	-2.3	-2.1	-1.4
G/C	-2.1	-2.0	-2.9	-1.8	-1.9	-1.2
U/A	-0.9	-1.7	-2.1	-0.9	-1.0	-0.5
G/U	-0.5	-1.2	-1.4	-0.8	-0.4	-0.2
U/G	-1.0	-1.9	-2.1	-1.1	-1.5	-0.4

Note that the value depends on the nucleotides at positions i and j (rows) and positions $i + 1$ and $j - 1$ (columns).

Predicted free-energy values (kcal/mol at 37^oC) for features of predicted RNA secondary structures, by size of loop:

size	internal loop	bulge	hairpin
1	-	3.9	-
2	4.1	3.1	-
3	5.1	3.5	4.1
4	4.9	4.2	4.9
5	5.3	4.8	4.4
10	6.3	5.5	5.3
15	6.7	6.0	5.8
20	7.0	6.3	6.1
25	7.2	6.5	6.3
30	7.4	6.7	6.5

This table provides values for the functions *ebi* (used for both internal and bulge loops) and *eh* (hairpin loops).

Recursion

There is the main recursion of the Zuker algorithm:

For all i, j with $1 \leq i < j \leq L$:

$$W(i,j) = \min \begin{cases} W(i+1,j) & (a) \\ W(i,j-1) & (b) \\ V(i,j) & (c) \\ \min_{i<k<j}\{W(i,k) + W(k+1,j)\}, & (d) \end{cases}$$

The equation considers the four cases in which (a) i is unpaired, (b) j is unpaired, (c) i and j are paired to each other and (d) i and j are not paired and encompass two smaller structures. In case (c) we reference the auxiliary matrix V.

The minimum folding energy E_{min} is given by $W(1, L)$.

To complete the main recursion, here is the computation of V:

$$V(i,j) = \min \begin{cases} eh(i,j) & (e) \\ es(i,j) + V(i+1,j-1) & (f) \\ \min_{i<i'<j'<j}\{ebi(i,j,i',j') + V(i',j')\} & (g) \\ \min_{i+1<k<j-1}\{W(i+1,k) + W(k+1,j-1)\} + a. & (h) \end{cases}$$

This part of the recursion considers the different situations that arise when bases i and j are paired, closing (e) a hairpin loop, (f) a stacked pair, (g) a bulge interior loop, or (h) a multi-loop.

Case (g) takes into account all possible ways to define a bulge or interior loop that involves a base pair (i', j') and is closed by (i, j). In each situation, we have a contribution from the bulge or interior loop and a contribution from the structure that is on the opposite side of (i', j').

Case (h) considers the different ways to obtain a multi-loop from two smaller structures and adds a constant contribution of a to close the loop.

Amino acids

Amino acids are small molecules that are the building blocks of proteins. Proteins serve as structural support inside the cell and they perform many vital chemical reactions. Each protein is a molecule made up of different combinations of 20 types of smaller, simpler amino acids. Protein molecules are long chains of amino acids that are folded into a three-dimensional shape.

Chemically, an amino acid is a molecule that has a carboxylic acid group and an amine group that are each attached to a carbon atom called the α carbon. Each of the 20 amino acids has a specific side chain, known as an R group, that is also attached to the α carbon. The R groups have a variety of shapes, sizes, charges, and reactivities. This allows amino acids to be grouped into 3 ctegories according to the chemical properties of their side chains.

Hydrophobic amino acids

They have side chain composed of carbon and hydrogen that are unlikely to form hydrogen bond with water.eg phenylalanine, and valine

Polar amino acids

They contain oxygen and/ or nitrogen in their side chain form hydrogeon bond with water.ex. serine, threonine, and asparagines

Charged Amino acids

Carry a + or – charge at biological pH.exa.glutamate ,Lysine

Amino acid Structure:

There are basically 20 standard amino acids having different structures in their side chains(R groups) . The common amino acids are known as α-amino acids because they have a primary amino group(-NH2) and a carboxylic acid group(-COOH) as substitutes of the a carbon atoms. Proline is an exception because it has a secondary amino group (-NH-), for uniformity it is also treated as alpha-amino acid.

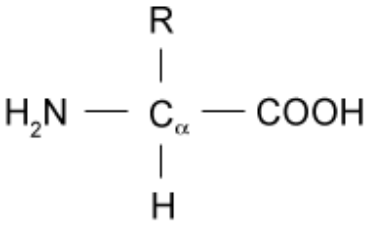


Fig.general structure of α-amino acids

General properties

The amino and carboxylic acid groups of amino acids readily ionize. At a pH(~7.4), the amino groups are protonated and the carboxyl acid groups are in their conjugate base(carboxylate) form, this shows that an amino acid that can act as an Acid and also a base. Amino acids can bear charged groups of opposite polarity, hence they are know as zwitterions or dipolar ions. The ionic property of the side chains influences the physical and chemical property of free amino acids and amino acids in proteins.

Polypeptide composition

- In 1902, Emil Fischer proposed that proteins are long chains of amino acids joined by peptide bonds

- Peptide bond: the special name given to the amide bond between the a-carboxyl group of one amino acid and the a-amino group of another

Peptides

peptide: the name given to a short polymer of amino acids joined by peptide bonds; they are classified by the number of amino acids in the chain

dipeptide: a molecule containing two amino acids joined by a peptide bond

tripeptide: a molecule containing three amino acids joined by peptide bonds

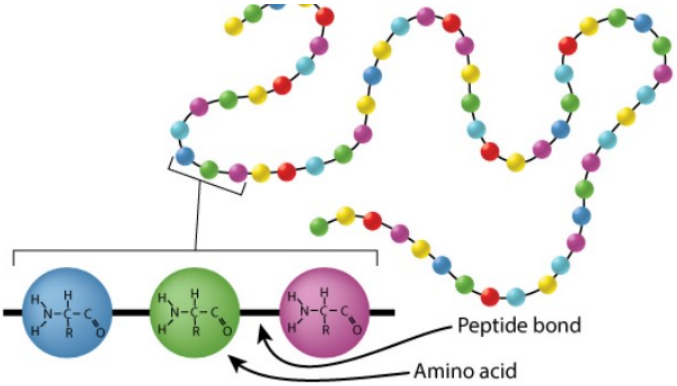
polypeptide: a macromolecule containing many amino acids joined by peptide bonds

protein: a biological macromolecule of molecular weight 5000 g/mol or greater, consisting of one or more polypeptide chains

Protein Structures

primary structure

The **primary structure** of a protein is the particular **sequence** of amino acids that form the backbone of a peptide chain or protein



Secondary Structure

Stretches or strands of proteins or peptides have distinct ***secondary structure***, dependent on hydrogen bonding. The two main types of secondary structure are the α-helix and the ß-sheet.

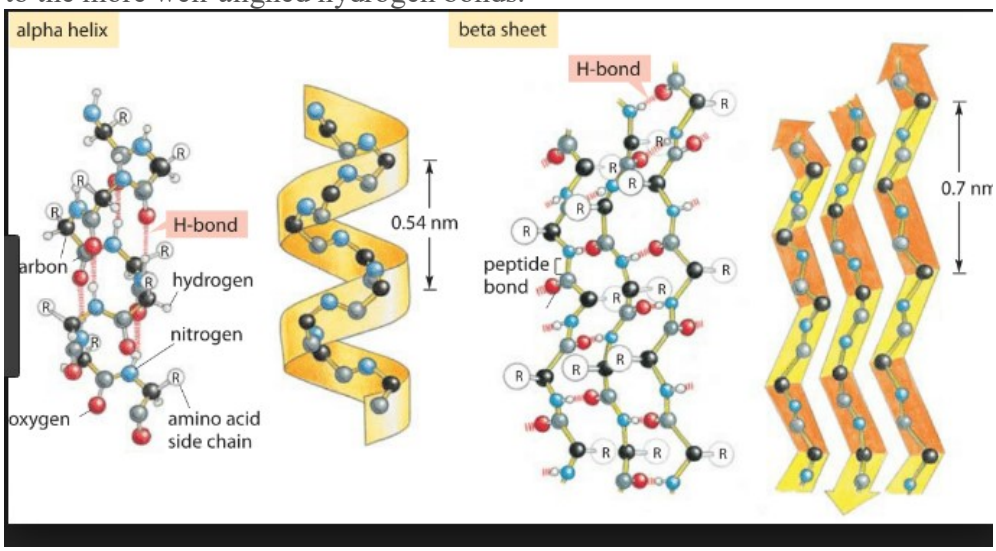
α-helix

The *α-helix* is a right-handed coiled strand. The side-chain substituents of the amino acid groups in an α-helix extend to the outside. Hydrogen bonds form between the oxygen of the C=O of each peptide bond in the strand and the hydrogen of the N-H group of the peptide bond four

amino acids below it in the helix. The hydrogen bonds make this structure especially stable. The side-chain substituents of the amino acids fit in beside the N-H groups.

ß-sheet

The hydrogen bonding in a *ß-sheet* is between strands (inter-strand) rather than within strands (intra-strand). The sheet conformation consists of pairs of strands lying side-by-side. The carbonyl oxygens in one strand hydrogen bond with the amino hydrogens of the adjacent strand. The two strands can be either parallel or anti-parallel depending on whether the strand directions (N-terminus to C-terminus) are the same or opposite. The anti-parallel ß-sheet is more stable due to the more well-aligned hydrogen bonds.



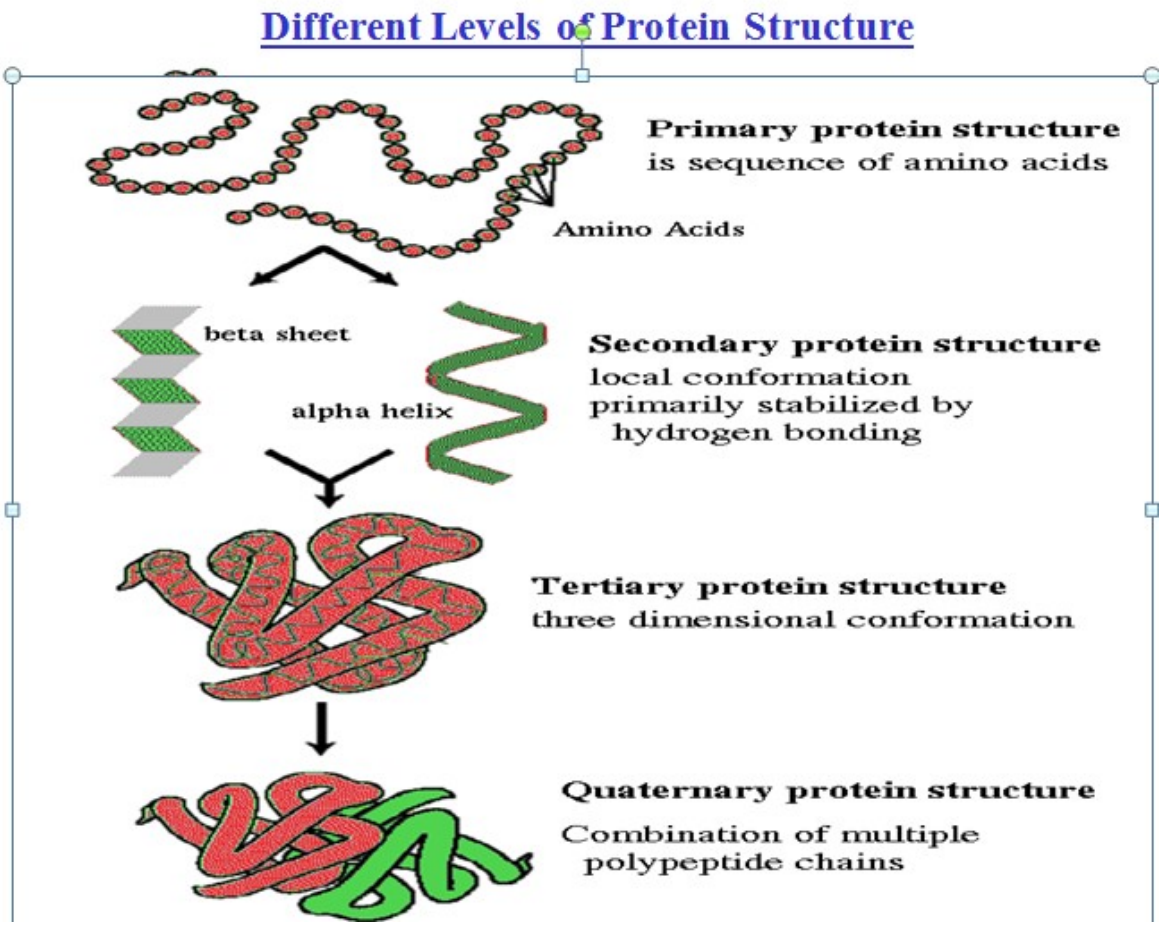
Tertiary Structure

The overall three-dimensional shape of an entire protein molecule is the ***tertiary structure***. The protein molecule will bend and twist in such a way as to achieve maximum stability or lowest energy state. Although the three-dimensional shape of a protein may seem irregular and random, it is fashioned by many stabilizing forces due to bonding interactions between the side-chain groups of the amino acids.

Quaternary Structure

Many proteins are made up of multiple polypeptide chains, often referred to as *protein subunits*. These subunits may be the same (as in a homodimer) or different (as in a heterodimer). The ***quaternary structure*** refers to how these protein subunits interact with each other and arrange themselves to form a larger aggregate protein complex. The final shape of the protein

complex is once again stabilized by various interactions, including hydrogen-bonding, disulfide-bridges and salt bridges. The four levels of protein structure are shown in Figure 2.



Protein folding

Proteins consists of linear chains of amino acids, but they do not simply flop in your cells, instead, the regarding proteins "fold" up into a particular three-dimensional conformation in solution (tertiary structure) , and this conformation helps the proteins to carry out the functions, which are responsible for.

Structure Prediction

It predicts the 3 dimensional shape of a protein from a given set of amino acid sequence.

Importance of protein structure

- knowledge of the structure of a protein enable us to understand its function and functional mechanism
- design better mutagenesis experiments
- structure-based rational drug design

Methods of protein structure prediction

Experimental methods

1.**X-ray Crystallography :**

- Determining protein structure directly is difficult
- X ray diffraction -studies must first be able to crystallize the protein and then calculate its structure by the way it disperses X - rays.

Problems:

- Only a small number of proteins can be made to form crystals.
- Very time consuming

2.**NMR**

- Use nuclear magnetic resonance to predict distances between different functional groups in a protein in solution.

- Calculate possible structures using these distances.

Problems:

- Not all proteins are found in solution.
- This method generally looks at isolated proteins rather than protein complexes.
- Very time consuming

Computational Methods

There are four major classes of algorithms for the prediction of proteins structure.

1. **ab initio**
   - Ab initio- or de novo- protein modelling methods seek to build three-dimensional protein models "from scratch", i.e., based on physical principles rather than (directly) on previously solved structures.
   - There are many possible procedures that either attempt to mimic protein folding or apply some stochastic method to search possible solutions (i.e., global optimization of a suitable energy function). These procedures tend to require vast computational resources, and have thus only been carried out for tiny proteins.
   - Since these techniques start with less information, their performance is not as good
   - To predict protein structure for larger proteins will require better algorithms and larger computational resources likethose afforded by either powerful supercomputer
2. **Homology based methods**

- Also known as comparative modellig
- Homology based methods work from the assumption that proteins with shared ancestry will have mutually conserved sequence and structure.
- The objective is to identify homologous proteins with known structures and to use these similar structures to predict the structure for an unknown protein.
- This is done using template elements analogous to building blocks, such as Legos. Correspondence information is derived from primary structure (i.e. sequence) similarity.

- Accordingly, homology based algorithms are the most reliant of the four classes, generally requiring a database of known proteins with at least 30% sequence similarity and that provide at least 90% template coverage.

- Since homology algorithms work closely with similar, known natural structures, they also tend to have the best accuracy, producing predictions with an RMSD on the order of 1 - 3Å.

- They also tend to be the fastest and easiest to implement with running times on the order of seconds.

Steps in comparative modeling

1. Identify a set of protein structures related to target proteins
   - Sequence database tools like BLAST and FASTA are used to identify related structures
2. Align the sequence of the target with the sequence of template proteins
   - Multiple sequence alignment tools like CLUSTAL W is used to find an alignment
3. Construct the model

-need to obtain the conserved regions.Backbone of the template structure is then aligned to these conserved region.

4. Model the loops
5. Model the side chains
6. Evaluate the model
   - Based on software packages like PROCHECK,WHATCHECK,VERIFY - 3D etc...

3. Threading:reverse protein folding

- New way of fold recognition
- Sequence is tried to fit in known structures
- Motif recognition
- Loop & Side chain modelling
- Fail in absence of known example
- is paid for with losses in accuracy and computation time

steps

- Given:
  - sequence of protein P with unknown structure
  - Database of known folds
- Find:
  - Most passible fold for P
  - Evaluate quality of such arrangement
- Places the residues of unknown P along the backbone of a known structure and determines stability of side chains in that arrangement

4 Fold recognition

- used to assign tertiary structures to protein sequences, even in the absence of clear homology.
- The ongoing development of such methods has had a significant impact on structural biology, providing us with an increasing ability to accurately model 3D protein structures using very evolutionary distant fold templates