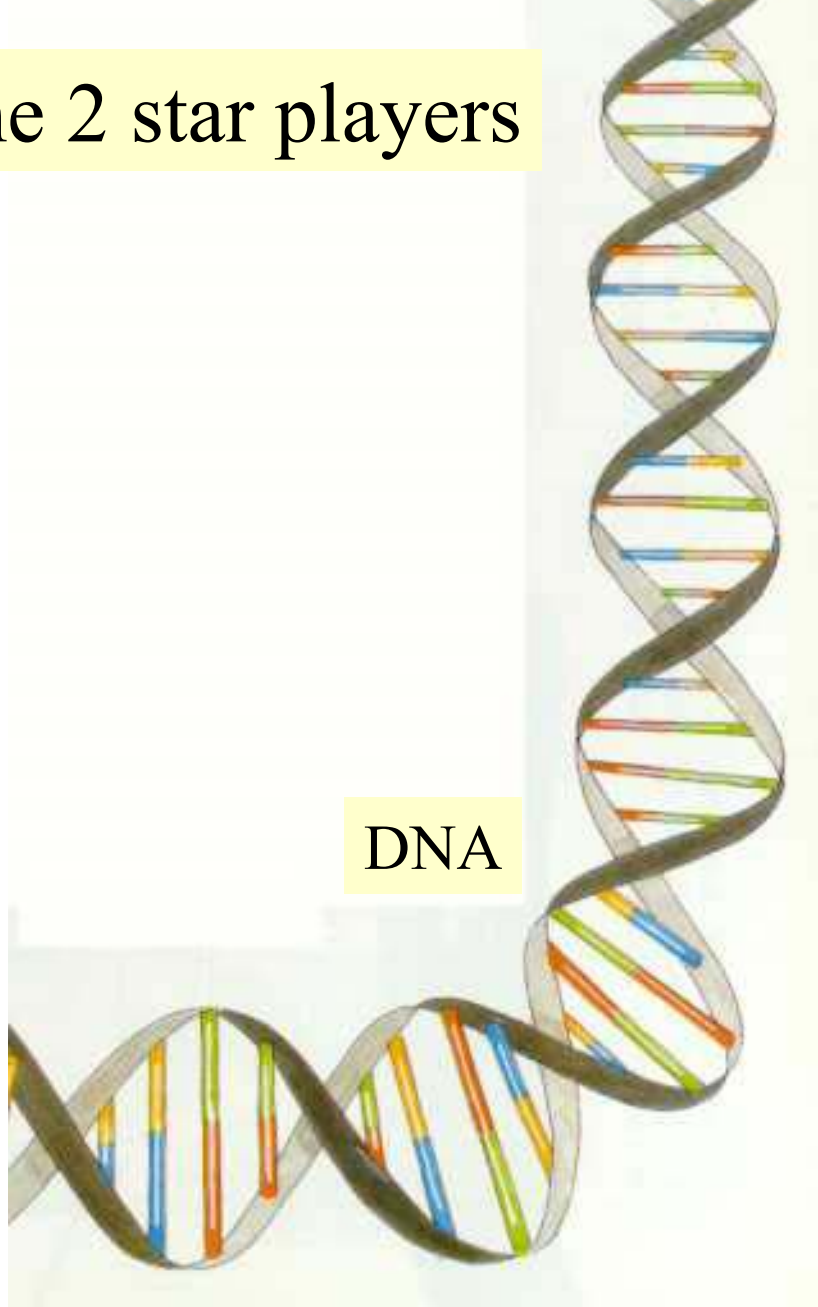
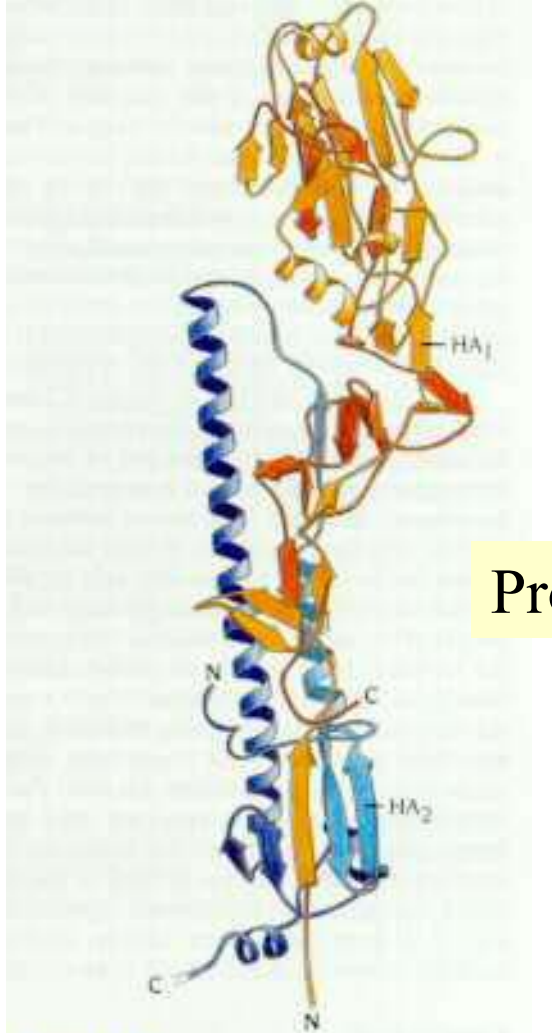
The 2 star players



DNA



Protein

Figure 8.21 Schematic diagram of the subunit structure of hemagglutinin from influenza virus. The structure comprises about 550 amino acids arranged in two chains HA_1 (red) and HA_2 (blue). The first half of each chain has a lighter color in the diagram. The subunit is very elongated with a long stemlike region built up by residues from both chains and includes one of the longest α helices known in a globular structure, about 75 Å long. The globular head is formed by residues only from HA_1. (Courtesy of Don Wiley, Harvard University.)

The Players

DNA

String with alphabet {A, C, G, T}

Nucleotides/Bases

RNA

String with alphabet {A, C, G, U} **Bases**

Protein

String with 20-letter alphabet

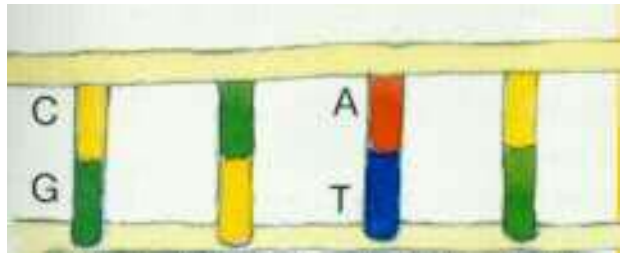
Amino acids/Residues

Central Dogma

- DNA acts as a template to replicate itself.
- DNA is trasnscribed into RNA.
- RNA is translated into **Protein**.

DNA

Complementary Bases



Proteins – Amino acids

amino acid	3 letter code	1 letter code
alanine	Ala	A
arginine	Arg	R
aspartic acid	Asp	D
asparginine	Asn	N
cysteine	Cys	C
glutamic acid	Glu	E
glutamine	Gln	Q
glycine	Gly	G
histine	His	H
isoleucine	Ile	I
leucine	Leu	L
lysine	Lys	K
methionine	Met	M
phenylalanine	Phe	F
proline	Pro	P
serine	Ser	S
threonine	Thr	T
tryptophan	Trp	W
tyrosine	Tyr	Y
valine	Val	V

Table 1.1: *Amino acid abbreviations*

Protein: Structure

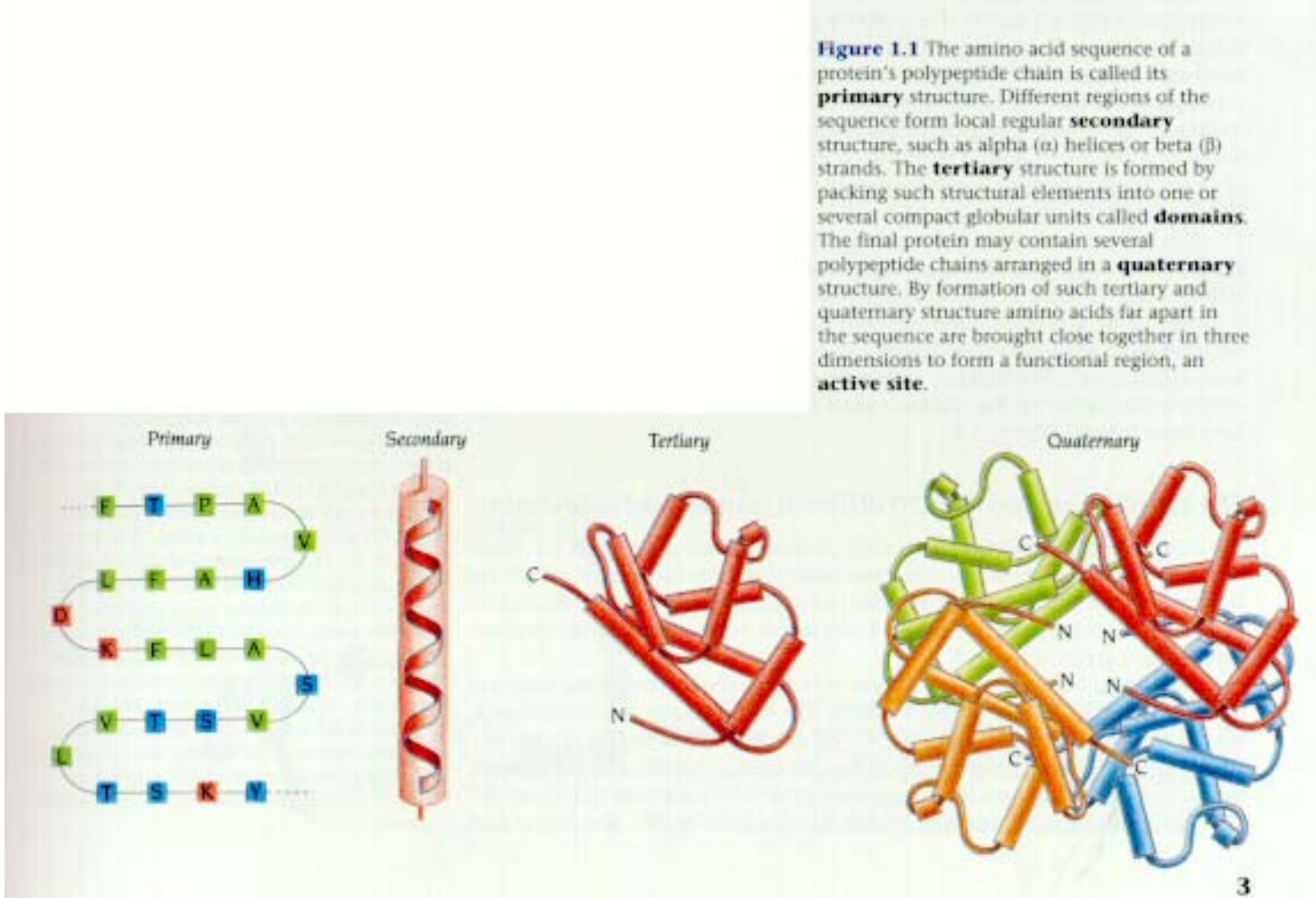


Figure 1.1 The amino acid sequence of a protein's polypeptide chain is called its **primary** structure. Different regions of the sequence form local regular **secondary** structure, such as alpha (α) helices or beta (β) strands. The **tertiary** structure is formed by packing such structural elements into one or several compact globular units called **domains**. The final protein may contain several polypeptide chains arranged in a **quaternary** structure. By formation of such tertiary and quaternary structure amino acids far apart in the sequence are brought close together in three dimensions to form a functional region, an **active site**.

Proteins: Active Sites

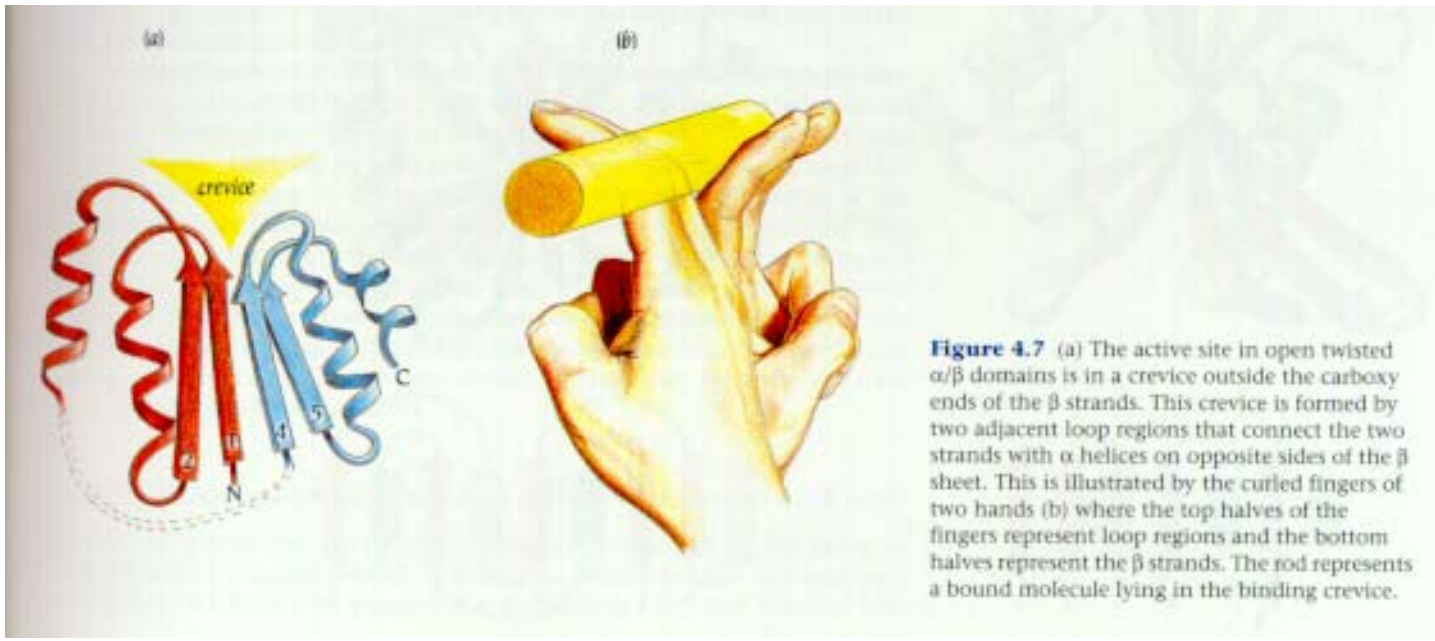
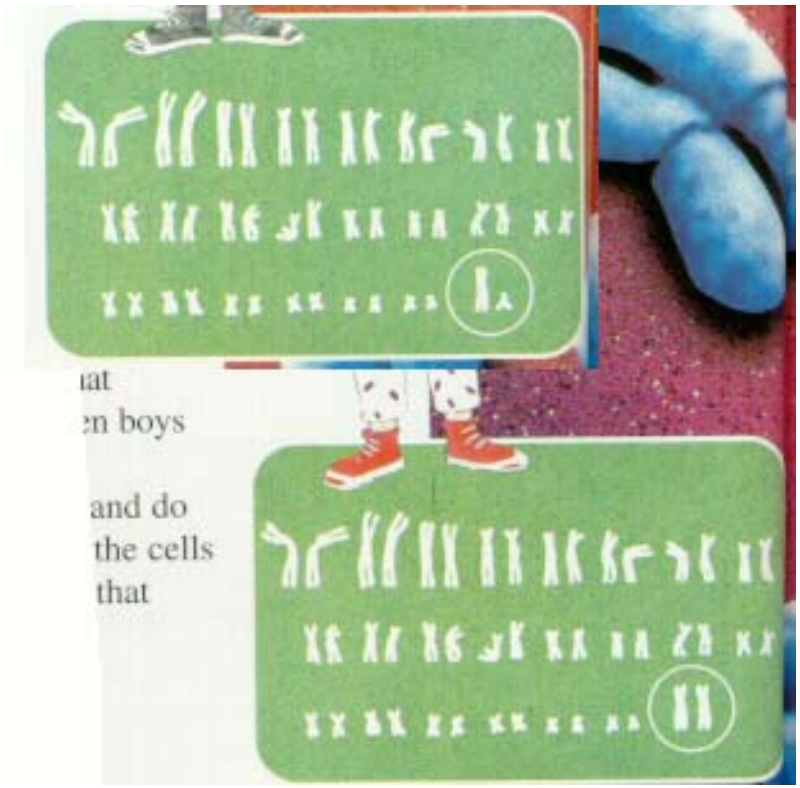
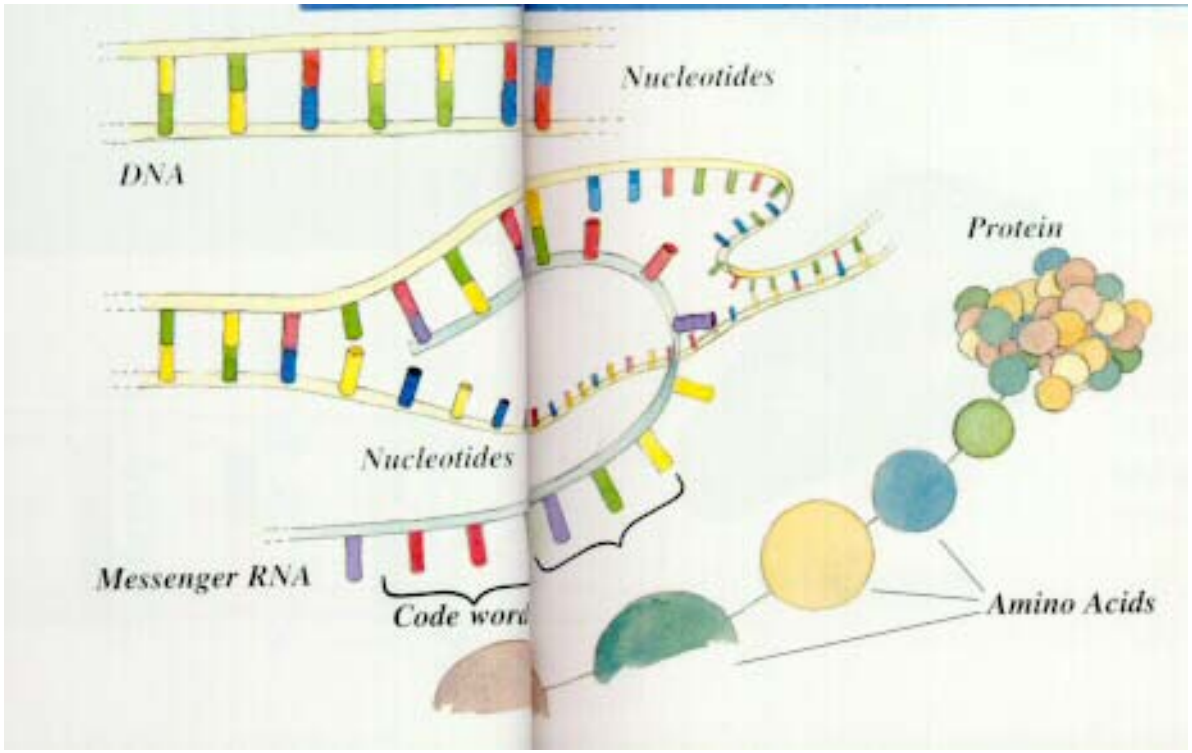


Figure 4.7 (a) The active site in open twisted α/β domains is in a crevice outside the carboxy ends of the β strands. This crevice is formed by two adjacent loop regions that connect the two strands with α helices on opposite sides of the β sheet. This is illustrated by the curled fingers of two hands (b) where the top halves of the fingers represent loop regions and the bottom halves represent the β strands. The rod represents a bound molecule lying in the binding crevice.

Chromosomes



DNA → RNA → Protein



RNA

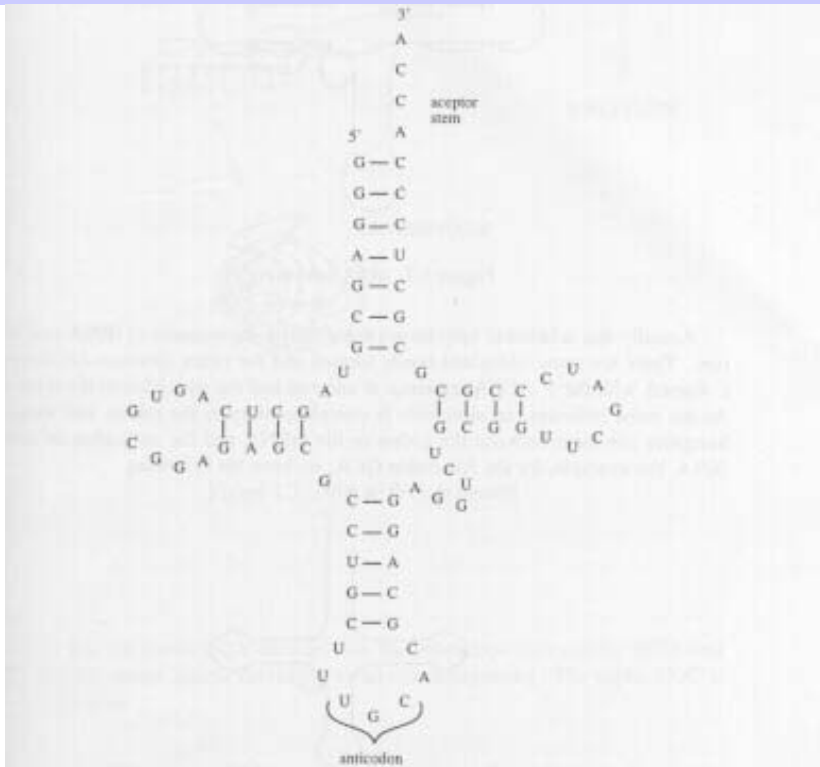


Figure 1.1: E. coli *Ala tRNA*

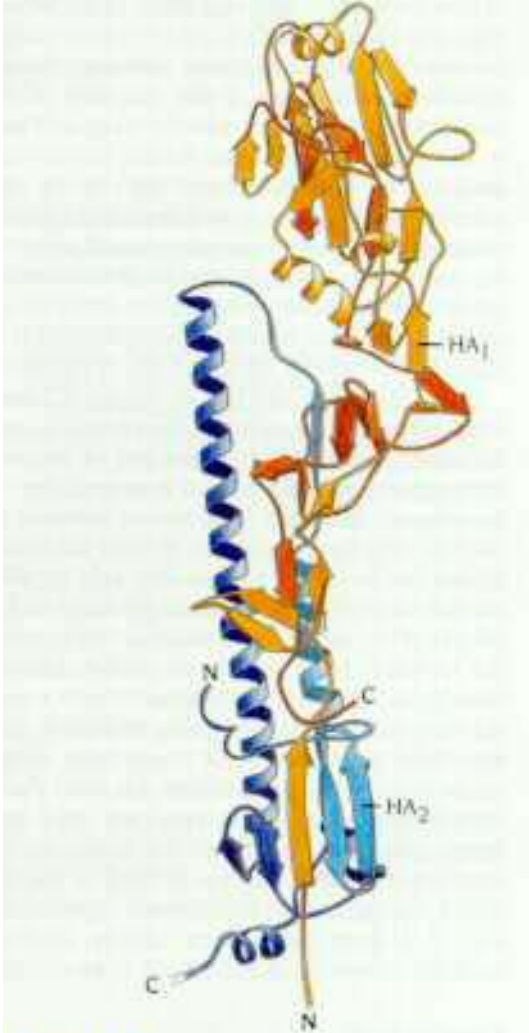
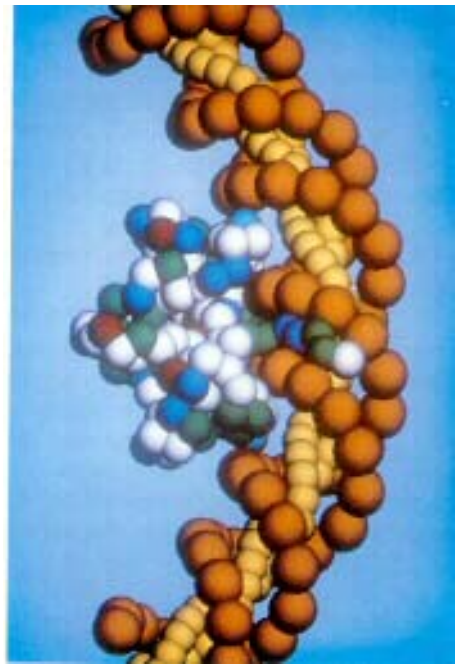


Figure 8.21 Schematic diagram of the subunit structure of hemagglutinin from influenza virus. The structure comprises about 550 amino acids arranged in two chains HA_1 (red) and HA_2 (blue). The first half of each chain has a lighter color in the diagram. The subunit is very elongated with a long stemlike region built up by residues from both chains and includes one of the longest α helices known in a globular structure, about 75 Å long. The globular head is formed by residues only from HA_1. (Courtesy of Don Wiley, Harvard University.)



String Matching Problem

(Approximate) String Matching

Input: Text **T** , Pattern **P**

Question(s):

- Does **P** occur in **T**?
- Find one occurrence of **P** in **T**.
- Find all occurrences of **P** in **T**.
- Count # of occurrences of **P** in **T**.
- Find longest substring of **P** in **T**.
- Find closest substring of **P** in **T**.
- Locate direct repeats of **P** in **T**.
- *Many More variants*

Applications:

- Is **P** already in the database **T**?
- Locate **P** in **T**.
- Can **P** be used as a primer for **T**?
- Is **P** homologous to anything in **T**?
- Has **P** been contaminated by **T**?
- Is *prefix*(**P**) = *suffix*(**T**)?
- Locate tandem repeats of **P** in **T**.

Input:	Text **T**; Pattern **P**
Output:	All occurrences of **P** in **T**.

Methods:

- Naïve Method **O(mn)** *time*
- Rabin-Karp Method **O(mn)** *time;* Fast on average.
- FSA-based method **O(n+mA)** *time*
- Knuth-Morris-Pratt algorithm **O(n+m)** *time*
- Boyer-Moore **O(mn)** time; Very fast on average.
- Suffix Tree method; **O(m+n)** *time*
- Shift-And method; Fast on average; Bit operations.