Protein sequence analysis

Aims:

1- Handling *sequence analysis server ExPASy Bioinformatics resource portal.*
2- Handling sequence alignment using VMD software.

Introduction

Proteins are the most abundant biological macromolecules, occurring in cells and parts of the cell. Proteins exhibit diversity of biological functions and are considered the most important final products of information pathways. They are the molecular instruments through which genetic information is expressed. They are constructed from the same set of 21 amino acids, covalently linked in a linear sequence.

Amino Acids

Proteins are polymers of amino acids, each amino acid is joined to its neighbor by loss of water and a specific covalent bond forms.
Common amino acids have an amino group and a carboxyl group bonded to the same amino acid (α carbon atom) [Fig. 1]. The Different amino acids have different R groups which vary in structure, size, and electric charge. The common amino acids are assigned three-letter abbreviations and one-letter code. Amino acids can be classified by its R groups [Fig. 2].

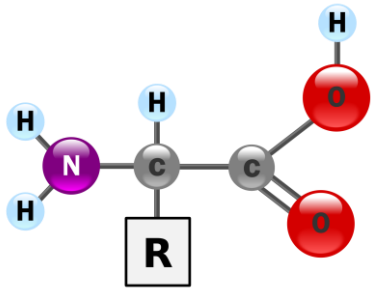

Figure 1 Amino acid

Peptides and Proteins

Peptides and proteins are polymers of amino acids. When few amino acids are joined, the structure is called *oligopeptide*. When many amino acids are joined, the product is called *polypeptide*. Proteins may have thousands of amino acid residues.

In polypeptides, the terminal with a free α-amino group is called amino terminal (or N-terminal) and that with a free carboxyl group is the carboxyl-terminal (C-terminal).

Proteins consisting of more than one polypeptide chain are called *multisubunit* proteins. If at least two are identical the protein is said to be *oligomeric*. Hemoglobin is an example [Fig. 3], it consists of four subunits two are identical and called α chains and the other two are called β chains. They are held together by noncovalent interactions.

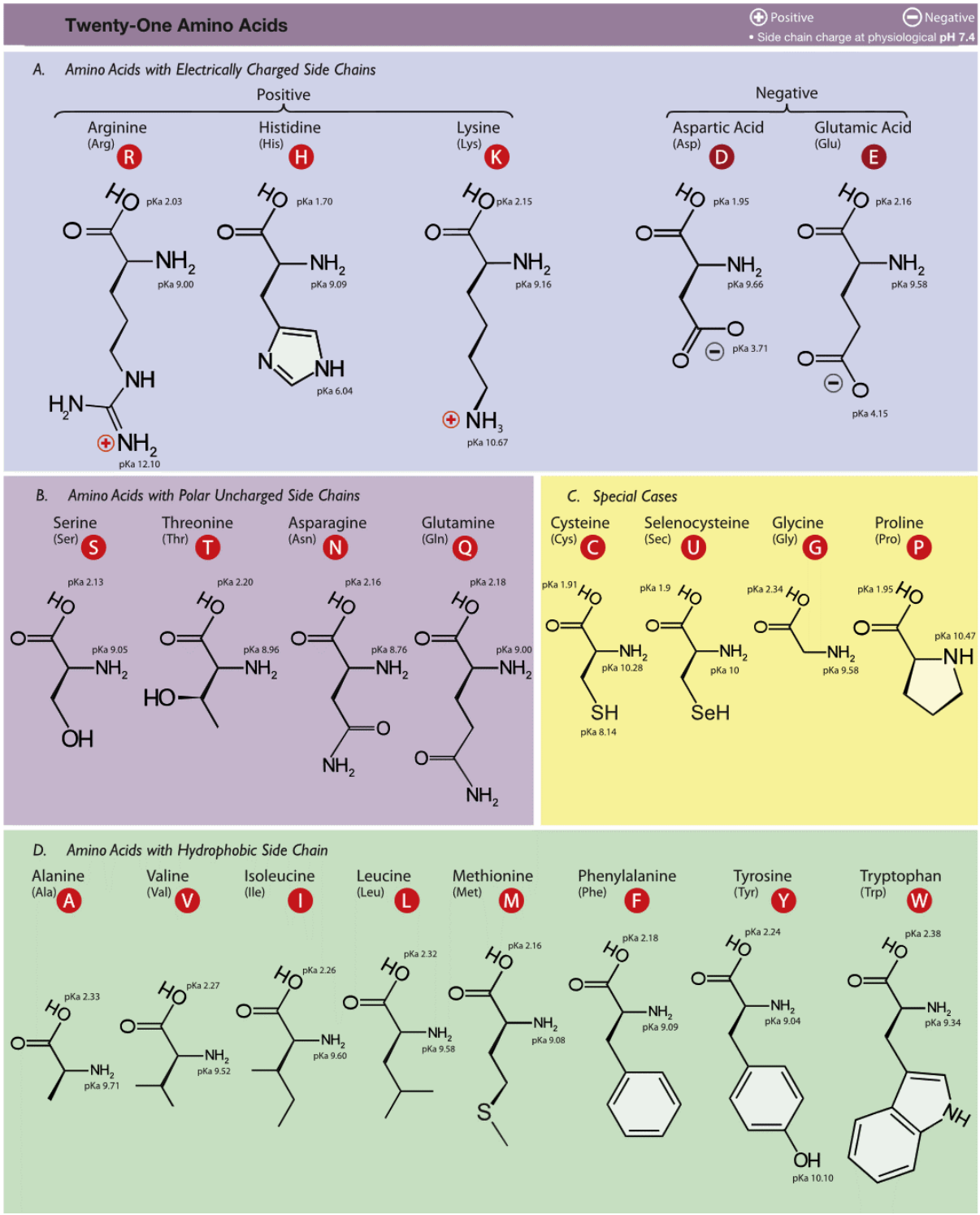

Figure 2 Classification of amino acids according to R groups.

Many proteins, for example enzyme ribonuclease A and chymotrypsin, contain only amino acid residues and no other chemical constituents. However, other proteins, called *conjugated proteins*, contain chemical constituents in addition to amino acids. The non-amino acid part of the conjugated proteins is called *prosthetic group*. As examples, lipoproteins contain *lipids*, glycoproteins contain *sugar* groups and metalloproteins contain a specific *metal*.

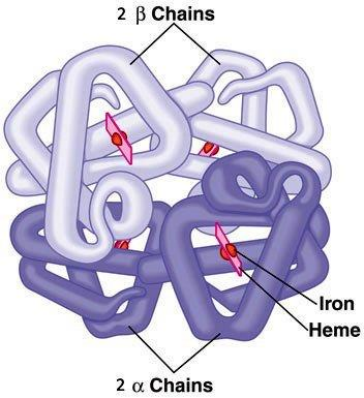

Figure 3 Hemoglobin molecule.

Levels of protein structure:

There are four different levels of protein structure. *Primary structure* refers to the amino acid sequence. *Secondary structure* refers to the arrangement of amino acid residues giving rise to specific structural pattern; it depends mainly on hydrogen bonding. *Tertiary structure* refers to the three dimensional folding of polypeptides. When protein has two or more polypeptides subunits, their arrangement in space is referred to as *quaternary structure*.

Theory

Sequence conservation

It is the amino acid(s) that is conserved through sequences under study. Proteins have regions of conserved domains that don't change in different organisms (Ex. is the GDD motif of the RNA polymerase).

Pair-wise percentage sequence identity

It is the percent identity between pair of sequence under study. They are different algorithms to get the percentage identity. For sequence identity of more than 30 % the protein fold may be the same.

Protein fold

It is the folding of amino acids to a desired shape. There are some distinct folds that mostly occur in proteins. The more predominate shapes are: Alpha helix and Beta sheets & turn. [Fig. 4]

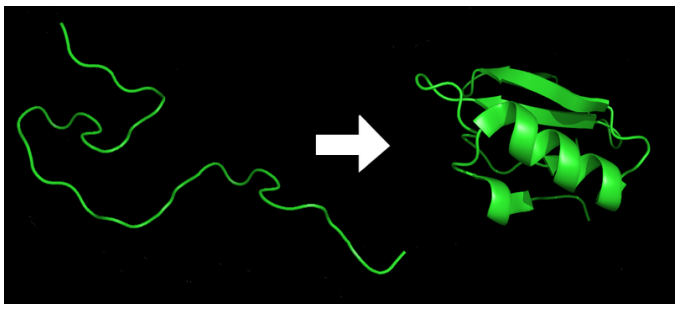

Figure 4 Protein folding.

FASTA

The FASTA file format includes the amino acid sequence in one-letter code, usually with 60 letters per line. (The most important is the sign ">", "larger than", on the first row. Alignment programs like CLUSTALW will use everything after the >-sign on that row as the title for the alignment).

Insertion

It is the case in which an amino acid(s) inserted between two amino acids in a sequence. As shown in *Fig. 5 line 2* an aspartic acid (D) is inserted between valine (V) and alanine (A)

Deletion

It is the case in which an amino acid(s) is absent in a sequence. [Fig. 5]

```
E S D I R T E E A I Y Q C C D L . P Q A R V . A I R S L T E R L
E S D I R T E E V I Y Q C C D L D P Q A R V D A I R S L T E R L
E S D I R T Z E A I Y Q C C D L D X Q A R V . A I K S L T E R L
```

Figure 5 Insertion and delation

Procedures:

1. Get the protein sequence for Hepatitis C Virus NS5b RdRp from different genotypes (1-6) from ExPASy proteomic server http://www.expasy.ch/ or National Center for Biotechnology Informatics (NCBI) web site http://www.ncbi.nlm.nih.gov/

2. Make sequence alignment of the downloaded sequences using ClustalW2 tool in the server of EMBL European Bioinformatics Institute http://www.ebi.ac.uk/Tools/msa/clustalw2/

3. Get the percent identity of the used sequences.

4. Repeat steps 2 and 3 but using VMD program (multiseq extension) instead of the ClustalW2 tool in EMBL server.

5. Using the following sequence for H1N1 flu virus, repeat the steps from 2 to 4.
   GLFGAMAGFIEGGWTGMIDGWYGYHHQNEQGSGYAADQKSTQNAI
   DGITNKVNSIIEKMNTQFTAVGKEFNNLERRIENLNKKVDDGFLDVW
   TYNAELLVLLENERTLDFHDSNVRNLYEKVKSQLRNNAKEIGNGCFE
   FYHKCDDECMESVKNGTYDYP

==========

Suggested web site:
http://proteinstructures.com/