CSC2427H Algorithm in Molecular Biology
Lecture #16 Protein Structure Overview

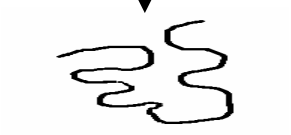
Lecturer: Michael Brudno
Scribe: Jifang Jiang
March 10th, 2006

1. Introduction

DNA

RNA

Transcription



Protein

Translation

Figure 1. The production of protein

RNA is transcribed from DNA, and then translates into protein.

Protein is a sequence of amino acids. It is built from 20 amino acids.

eg. MACILVGT…..*

It starts on M (MET, Methionine) since M is translation of start codon AUG.

* indicates the end of the sequence (Stop codon does not code any amino acid).

2. Distinct aspects of Protein Structures

2.1 Primary structure

The sequence of amino acid is considered as primary structure. Amino acids are joined together by peptide bond. The peptide linkages, along with the alpha-carbons to which side chain are attached, form the protein backbone.

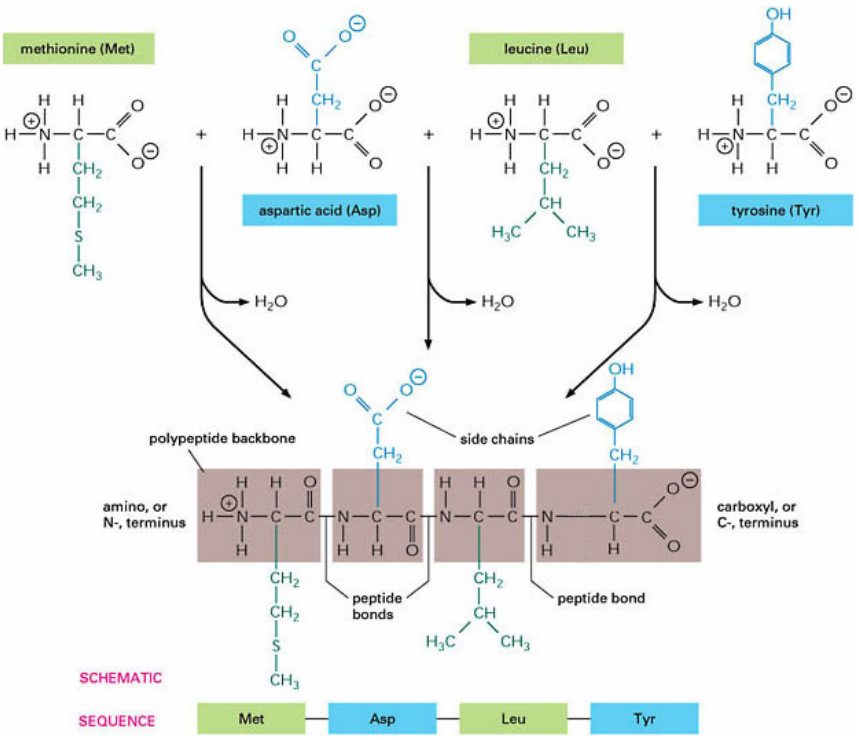
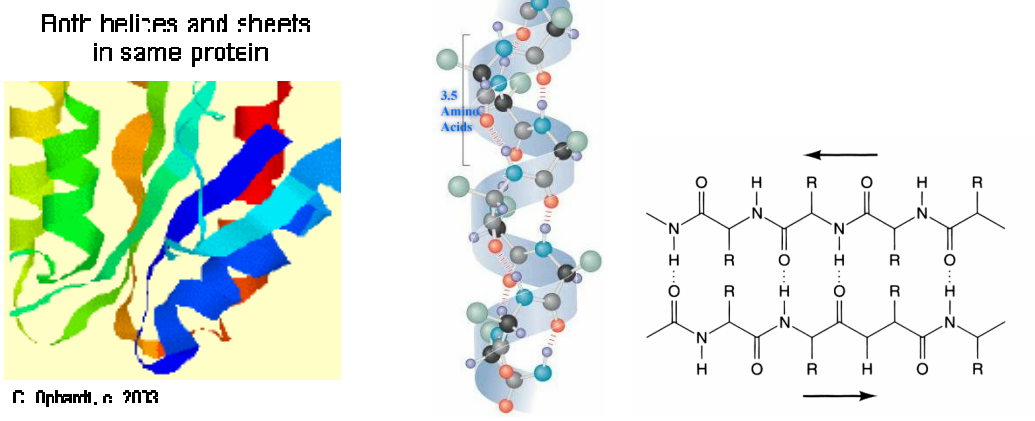


Figure 2. Protein Sequence

Each type of protein differs in its sequence and number of amino acids. The two ends of a polypeptide chain are chemically different: the end carrying the free amino group (NH3+, also written NH2) is the amino, or N-, terminus, and that carrying the free carboxyl group (COO-, also written COOH) is the carboxyl, or C-, terminus. The amino acid sequence of a protein is always presented in the N to C direction, reading from left to right. [1]

2.2 Secondary Structure

As being synthesized, most polypeptide chains fold up into a stable 3D structure. A few characteristic patterns appear frequently in folded structure. The most common substructures include: alpha-helices and beta-sheets.



(a) (b) (c)

Figure 3. Secondary Structures: (a) Both alpha-helices and beta-sheets structure (b) alpha-helix structure (c) beta-sheet structure Copyright (C) 2000,2001,2002 Free Software Foundation, Inc.

An alpha-helix is a 5.4Å wide corkscrew-type structure where the side chains project outward from the helix. Each amino acid results in a 100° turn in the helix, and corresponds to a translation of 1.5 Å along the axis. The amino acids chain has 3.6 amino acids per turn in alpha-helix. [3]. The helix is formed by hydrogen bonds between the CO group of each amino acid and the NH group of the residue four positions C-terminal (n = 4). Some amino acids have a greater tendency to be found in alpha-helix. Alanine, glutamine, leucine and methionine are commonly found in alpha-helices, whereas praline, glycine, tyrosine and serine usually are not[2].

The beta-sheet consists of two or more polypeptide chain that are arranged adjacently and in parallel within the same protein, but with alternating orientation such that hydrogen bound can form between the two strands. The amino acid chain is almost fully extended throughout the strand. The α-C atoms of adjacent strands stand 3.50 Å apart[4].

The structure of protein can be experimentally determined using MR crystallography, which is expensive and time consuming. Around 30,000-40,000 protein structures are known so far. There is a contest held every year called CASP, to see who can predict the structure of protein best.

2.3 Tertiary Structure

It is the three dimensional folding model of a protein. It describes the overall structure of various parts, including each atom position in protein. In the protein structure, an independently folded unit is called a protein domain. It is often joined by a flexible segment of the polypeptide chain.

The structure of protein is not rigid. Many factors can cause the orientation of backbone being changed, such as heat, PH, oxidants.

Many diseases are due to the protein can not fold correctly, such as mad-cow disease.

3. Structural Classification of Protein (SCOP) database

If two protein sequences share 25% residues, then they are considered coming from the same gene sequence, and their structure are almost identical. The proteins with same structure will have same function. Therefore, structure of a protein plays an important role in predicting the function of newly discovered protein.

The SCOP database is largely manual classification of proteins (One of SCOP mirror site: http://scop.berkeley.edu/index.html). It organizes proteins in a hierarchy way, from class down to fold, superfamily, and family. SCOP defines 10 classes on the top level based on similarities of their structure:

- All alpha proteins
- All beta proteins
- Alpha and beta proteins (a/b)
  Mainly parallel beta sheets (beta-alpha-beta units)
- Alpha and beta proteins (a+b)
  Mainly antiparallel beta sheets (segregated alpha and beta regions)
- Multi-domain proteins (alpha and beta)
  Folds consisting of two or more domains belonging to different classes
- Membrane and cell surface proteins and peptides
- Small proteins
  Usually dominated by metal ligand, heme, and /or disulfide bridges
- Peptides

- Designed proteins
- Non-protein structures

Actually, the last three are not true classes. There are many folds in each class. For example, Fold Globin-like is in the first class all alpha protein. It has a core of 6 helices; folded leaf, partly opened.

SCOP is a valuable resource of protein data structure. In SCOP database, proteins with similar sequences, or similar structures and functions are considered have a solid evolutionary link, and will be grouped as families, and share common ancestry. Therefore, SCOP has detailed evolutionary information. But since it is almost completely manually derived, some families or folds may not be as thoroughly detailed as others protein structural databases CATH and FSSP, which are a combination of manual and automated, and purely automated, respectively[5].

Reference:

[1] http://www.accessexcellence.org/RC/VL/GG/prot_Struct.html

[2] Andreas D. Baxevanis, Bioinformatics A Practical Guide to the Analysis of Genes and Proteins, 2nd Edition, P263

[3] http://en.wikipedia.org/wiki/Alpha_helix

[4] http://en.wikipedia.org/wiki/Beta_sheet

[5] Caroline Hadley and David T Jones, A systematic comparison of protein structure classifications: SCOP, CATH and FSSP, Structure. 7(9), 1999, 1099-1112.