Alma Mater Studiorum

Università di Bologna

Facoltà di Scienze Matematiche, Fisiche e Naturali

Dottorato di Ricerca in Biotecnologie Cellulari e Molecolari

Ciclo XX

FROM “WET BIOLOGY” TO STATISTICAL ANALYSIS OF STRUCTURAL FEATURES WITH BIOINFORMATICS TOOLS

Settore scientifico disciplinare di afferenza: BIO10

Tesi presentata da:
Paola Marani

Relatore:
Chiar.^{ma} Prof. Rita Casadio

Coordinatore:
Chiar.^{mo} Prof. Lanfranco Masotti

Esame finale anno 2008

Alle mie nonne

TABLE OF CONTENTS

Outer membrane Proteins in Gram-negative bacteria

INTRODUCTION

CHAPTER 1

- 1.1 Gram-negative bacteria: *Escherichia coli*
- 1.2 Biological Membranes: structure and function
- 1.3 Membrane proteins

CHAPTER 2

- 2.1 Identification of bacterial outer membrane proteins by computational approaches

CHAPTER 3

- 3.1 Secretion system in *Gram-negative* bacteria: autotransporter
- 3.2 Conclusion

References

Protein-DNA interactions

CHAPTER 4

- 4.1 Biological Macromolecules: Protein and Acid nucleic
- 4.2 Database: PDB
- 4.3 DNA binding proteins
- 4.4 Enzyme classification: EC number

- 4.5 Structural classification of proteins

References

List of pubblications

Schools attended

Attività didattica: tutor

International Schools: local committee

PAPER

INTRODUCTION

All living organisms are composed of cells and can be unicellular (only one cell like bacteria) or multicellular (higher organisms are a complex assembly of many cells with different functions).

Based on the organization of their cellular structures, all the cells can be divided into two groups: prokaryotic and eukaryotic.

Prokaryotic cells, like bacteria, are generally much smaller and relative simple, than eukaryotic cells, and because of their small size they are able to be structurally more simple.

The surface area of a cell relative to the volume decreases as size of cell increases, this limits the size of the cell because cells must be able to exchange materials with their surroundings. So the smaller a cell, the greater is its surface-to-volume ratio, this means that nutrients can easily and rapidly reach any part of the cells interior.

The eukaryotic cell is larger and the limited surface area when compared to its volume means nutrients cannot rapidly diffuse to all interior parts of the cell. That is why eukaryotic cells require a variety of specialized internal organelles, that are compartments for special metabolic functions and they have developed transport mechanisms that may be necessary to support their larger size.

CHAPTER 1

1.1 Gram-negative bacteria: *Escherichia coli*

We can differentiate bacterial species into two large groups: Gram positive and Gram negative by a method developed by Hans Christian Gram (1853-1938). The Gram stain is used to classify bacteria on the basis of their forms, sizes, cellular morphologies, and Gram reactions; this classification is based on the chemical and physical properties of bacterial cell walls.

Escherichia coli is a Gram-negative bacterium that has been extensively used as a model organism for biological studies because of ease of manipulation. The name comes from its discoverer, Theodor Escherich (1857-1911).

E. coli is a rod shaped, facultative anaerobic, non-sporulating, enteric bacterium, which can be found in the lower intestinal tract of warm-blooded animals like humans (Figure 1).

Normally, *E. coli* does not cause disease although some strains frequently cause diarrhea in travelers, and it is the most common cause of urinary tract infections; only one strain, designated O157:H7, is particularly virulent and has been responsible for several dangerous outbreaks of infections in people by producing a toxin which damages the intestines.

An average *E. coli* cell is about 3 μm in length and about 1 μm in diameter (Madigan et al., 1997).

As in all Gram-negative bacteria, the *E. coli* cell envelope consists of:

- the inner membrane (also referred to as the cytoplasmic membrane or the plasma membrane)
- the peptidoglycan containing periplasm
- the outer membrane

(Madigan et al., 1997).

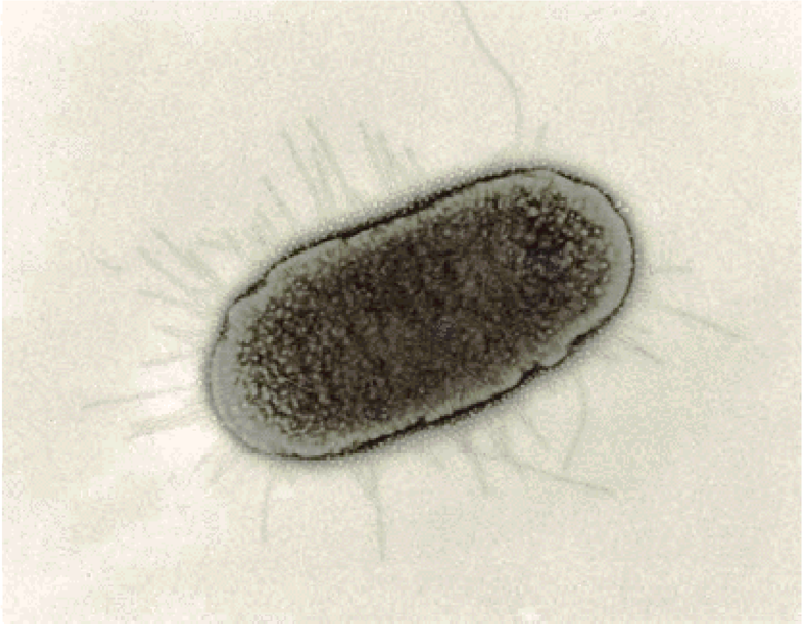


Figure1 A transmission electron micrograph of *Escherichia coli (E.coli)*, negatively stained to enhance contrast. Note the projecting pili, which may be involved in mechanisms of infection.

http://www.wadsworth.org/databank/ecoli.htm

1.2 Biological Membranes: structure and function

Cells are surrounded by a membrane, which contains a wide variety of biological molecules, primarily lipids, proteins and sugars, and separate the inside of the cell from the outside environment. In this way, a compartment is formed in which all the essential cellular processes can take place.

Membrane lipids are amphiphilic molecules, which means that they have a hydrophilic headgroup (water-loving, or polar) and a hydrophobic tail (water-fearing, or non-polar) that spontaneously assemble into a bilayer sheets because the hydrophobic tails are forced to stay together by water molecules. All the membranes have this common and general structure where proteins are embedded. The structure is highly fluid and most of the lipid and protein molecules can move about in the plane of the membrane (Singer and Nicolson, 1972).

The central hydrophobic region is about 20-30 Å (Angstroms) and the polar interface region on both sides of the hydrophobic region is about 10-15 Å thick (White and Wimley, 1999).

Biological membranes are not only permeability barriers between different compartments in living cells and between cells and their environment but can also be looked upon as compartments in themselves. Many biological processes take place in and across membranes.

Small non polar molecules are thought to be able to diffuse spontaneously across membranes allowing cells to communicate with each other and with the surrounding media, and in this way they can regulate many biochemical processes, while the transport of polar and larger molecules must be facilitated and sometimes needs an input of energy.

These functions are carried out by the proteins embedded in the lipid bilayer.

Membrane proteins regulate the transport of different ions and nutrients through channels and transporters, and transmit information in and out of the cell via receptors. Thus, membrane proteins are important for maintaining the function of a biological membrane and for the survival of the cell.

The lipid and protein molecules are held together mainly by non-covalent interactions, while sugars are attached by covalent bonds to some of the lipid and protein molecules and are found on one side of the membrane only.

The cell envelope of Gram-negative bacteria is composed of two distinct lipid membranes, the inner membrane (IM) and the outer membrane (OM). The two membranes have an entirely different structure and composition. Whereas the IM is a symmetric phospholipid bilayer, the OM is an asymmetrical bilayer, consisting of phospholipids in the inner leaflet and lipopolysaccharides (LPS) in the outer leaflet. Lipopolysaccharide has long carbohydrate chains that cover the cell surface and protects the cell from hydrophobic compounds and toxins. Additionally, these membranes differ with respect to the structure of the integral membrane proteins. Whereas integral IM proteins typically span the membrane by forming 15-20 amino acid long hydrophobic α-helices, connected by loops located outside the lipid bilayer, integral OM proteins (OMPs) generally consist of antiparallel amphipathic β-strands that fold into cylindrical β-barrels.

β-barrels have hydrophilic interior and hydrophobic residues pointing outward to face the membrane lipids.

The major determinant of the protein orientation in the membrane is stated in the ‘positive inside rule’: α-helices loops that are located at the cytoplasmic side are enriched in positive amino acids (von Heijne, 1989).

Both membranes also contain lipoproteins, which are anchored to the membranes via an N-terminal *N*-acyl-diacylglycerylcysteine, with the protein moiety usually facing the periplasm in the case of *Escherichia coli*.

The zwitterionic, non-bilayer prone lipid phosphatidylethanolamine (PE) is the major lipid in *E. coli* and constitutes 70-80% of all phospholipids in the cell (Dowhan, 1997).

The remaining phospholipids are phosphatidylglycerol (PG) and cardiolipin (CL) which constitute 20-25% and 5% or less of the total phospholipid pool respectively.

The periplasm contains a peptidoglycan layer (also called murein) that is covalently attached to the outer membrane by the murein lipoprotein. Peptidoglycan consists of alternating N-acetyl glucosamine and N-acetyl muramic acid forming linear chains that are linked together by peptide bridges (Madigan et al., 1997).

The OM functions as a selective barrier that protects the bacteria from harmful compounds, such as antibiotics, in the environment and the proteins here located function, for example, as specific pores for nutrient uptake (Schulz, 2002).

Unlike the IM, the OM is not energized by a proton gradient and ATP is not available in the periplasm. In the absence of readily available energy sources, nutrients usually pass the OM by passive diffusion via an abundant class of trimeric OMPs called porins.

Porins form water-filled channels that allow the passage of small hydrophilic solutes with molecular weights up to ~600 Da.

All the components of the OM are synthesized in the cytoplasm or at the cytoplasmic face of the IM, and they have to be transported across the IM and through the periplasm to reach their destination and to assemble into the OM.

1.3 Membrane proteins

Membrane proteins can be classified into several subgroups, depending on their mode of interaction with the membrane; integral membrane proteins, peripheral and interfacial (monotopic) proteins.

Peripheral membrane proteins are associated with the surface of the membrane trough electrostatic and/or hydrophobic interactions or through interactions with integral membrane proteins.

The interfacial membrane proteins consist of one hydrophobic part that anchors the protein into the interface and one hydrophilic part that is soluble.

Both peripheral and interfacial membrane proteins can detached from the membranes in the presence of high salt or under alkaline conditions. In contrast, integral membrane proteins are resistant to high salt treatment, they transverse the membrane with one or more transmembrane (TM) segments by interacting directly with the hydrophobic interior of the bilayer.

Another type of membrane protein are the membrane-anchored proteins which are covalently attached via either fatty acids or a glycosyl-phosphatidylinositol (GPI) anchor.

Most transmembrane proteins appear to be the α-helical type and are found in all organisms. β-barrel type membrane proteins have only been found in the outer membranes of mitochondria, chloroplasts and *Gram-negative* bacteria (Figure 2). The two above conformations are the simplest arrangements for satisfying the potential H-bonds of the polypeptide backbone in the hydrophobic interior of the lipid bilayer. Generally transmembrane proteins are oriented perpendicular to the plane of the membrane and the TM segments are connected to each other by loops on either side of the membrane.

Integral membrane proteins are dynamic, stable structures optimized for their function.

This simple picture of integral membrane protein architecture came originally from 2-D crystals of bacteriorhodopsin, a light-driven protein pump from Halobacterium that contains seven transmembrane α-helices and was the first transmembrane protein shown to consist of a helix bundle (Henderson and Unwin, 1975).

With the increased number of high resolution structures this picture has changed, revealing more complex architectures: helices vary greatly in length and tilt angles relative to the membrane, some have functionally important half helices (short helices) as in the aquaporins where two half-helices point their N-terminal ends towards each other in the centre of the membrane forming the pore region (Stroud et al., 2003) or as in the pore helices in potassium channels (Jiang et al., 2003).

In many ion-channels an entire domain, the “voltage sensor domain” (VSD) is partially located in the interface part of the membrane (Jiang et al., 2003).

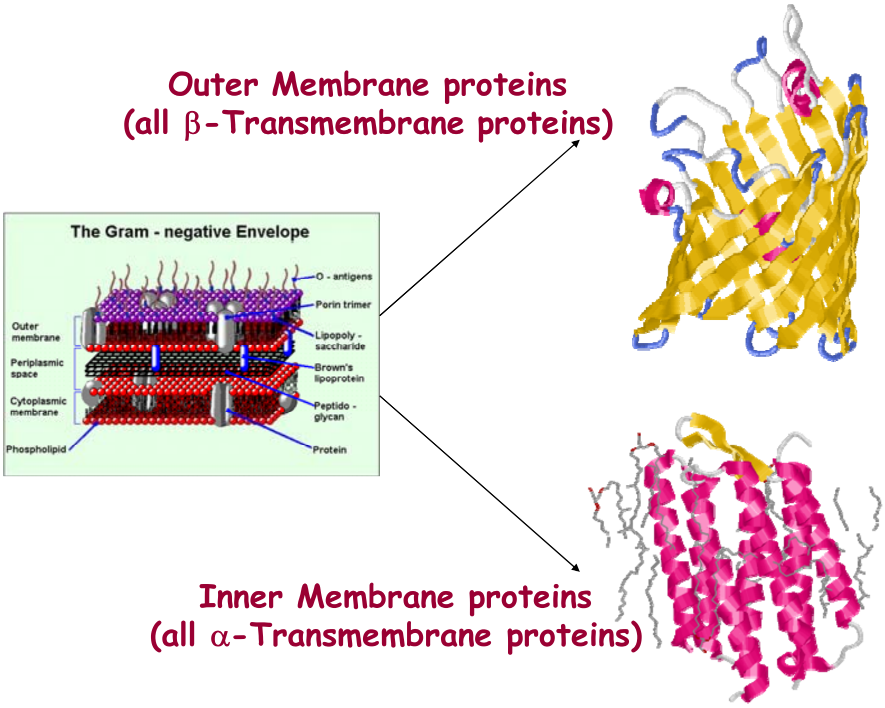


Figure 2 Data from known structures have shown that two basic architectural elements are found in transmembrane proteins: bundles of α-helices or amphipathic β strands forming a closed barrel.

CHAPTER 2

2.1 Identification of bacterial outer membrane proteins by computational approaches

In the post-"omic" era the analysis of the enormous amount of sequencing data is regarded as one of the major challenges faced by researchers.

Methods for large-scale analysis are necessary not only for a correct protein annotation, but also for focusing on specific protein categories or predicting their interaction. One focus of this data analysis is uncovering the genetic blue print of different organisms by automatically analyzing protein sequences for functional assignment and annotation by means of extensive homology search with PSI-BLAST or hidden Markov models. However there is still a substantial number of uncharacterized proteins, including hypothetical proteins (with homologues of unknown function) or unique proteins (without known homologues) that deserve further characterization.

Predictors of membrane protein structures can be used to filter genomes and find new membrane proteins without sequence homologues.

Due to experimental difficulties in expression and crystallization of membrane proteins a bioinformatic approach can be very important to guide the experimentalist toward targets that are highly likely to correspond to true instances of the particular kind of gene or protein of interest.

To this aim the Biocomputing Group (www.biocomp.unibo.it) has integrated a set of independent predictors that have been developed and have tested their discriminating capability on the genome sequenced of *Escherichia coli* K12 (Casadio et al., 2003).

The Biocomputing Group has implemented signal peptide predictor, and has developed two well performing predictors of the topography of inner all-alpha and outer all-beta membrane proteins, endowed with filters that minimize the rate of false positives (proteins falsely predicted in the category).

Hunter (Casadio et al., 2003) is a suite of these three predictors which are combined in an efficient manner.

All predictors use as input the sequence profile derived from multiple alignment of the target chain towards the non-redundant database.

Proteins have intrinsic signals that govern their transport and localization in the cell: a secretion hydrophobic marker, or signal peptide.

Sequences of outer membrane proteins have signal peptides: the secretion marker is also a marker of outer membrane proteins, situated in the N-terminal part of the protein sequence and from 20 to 30 amino acid long.

The implementation essentially reflects the general rules that can be derived from a statistical analysis of the protein chains from *E.coli* K12, that are well-annotated in Swiss-Prot:

- Globular proteins can be endowed or not with signal peptides
- Similarly all-alpha membrane proteins can include or not signal peptides
- All-beta outer membrane proteins contain signal peptides

It is evident that the presence of a signal peptide in the sequences is a characteristic tag of outer membrane proteins (Nielsen et al., 1997).

Furthermore not all the proteins with a signal peptide are outer membrane proteins.

The flow-chart of Hunter is shown in Figure 3.

The protein content of the genome is filtered with the signal peptide predictor. If the protein does not have a signal peptide it is then filtered with a neural-network-based predictor. A chain with transmembrane helixes predicted is classified as all-alpha membrane proteins, if not it is classified as globular protein.

If the protein has a signal peptide is sent to the filter for all-alpha membrane proteins and if retained, it is accordingly classified; if not it is then presented to the filter for all-beta-membrane proteins and if retained, it is accordingly classified; if not it is classified as globular.

'Hunter' has been used to filter the twilight zone of the genome of *E.coli* K12. Out of the 18 outer membrane proteins, 6 chains have no homologues in Swiss-Prot, 4 are annotated as hypothetical proteins, and the remaining have homologues.

But are these predictions reliable?

To answer at this question we decided to check the real localization in *E. coli* of these protein throughout experimental verification.

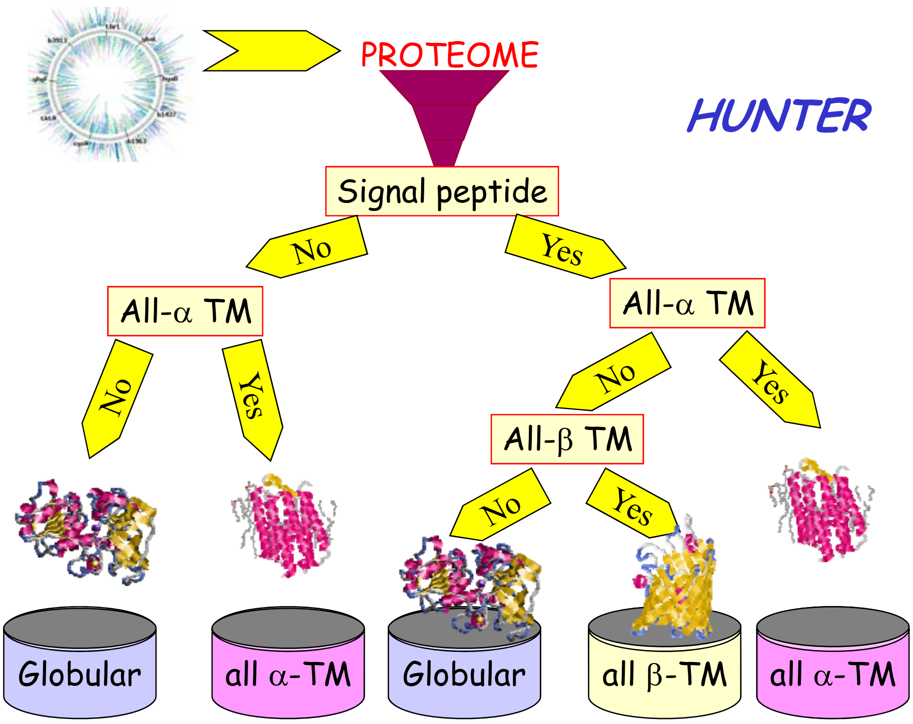


Figure 3 Flow chart of Hunter a suite of predictors (Casadio et al., 2003)

CHAPTER 3

3.1 Secretion system in *Gram-negative* bacteria: autotransporter

Secretion of proteins into the extracellular environment is important to almost all bacteria, since it mediates interactions between pathogenic or symbiotic bacteria with their eukaryotic hosts.

Autotransporters constitute the largest family of secreted proteins in *Gram-negative* bacteria, many are very large proteins, ranging in size from 90 to 200 kDa.

Autotransporter secretion is known to involve the insertion into the outer membrane of the conserved C-terminal beta barrel domain and the translocation across the outer membrane of the functional domain present at the mature (signal sequence cleaved) N-terminus.

One of our predicted protein, Swiss Prot code : YFAL 130 kDa, is supposted to be an autotransporter protein.

3.2 Conclusion

From the list of 18 new outer membrane proteins predicted by ‘Hunter’ (Casadio et al., 2003) predictor in the unannotated portion of the *E. coli* proteome, we initially chose 11 proteins, characterized by different lengths and different numbers of predicted β-strands, for experimental localization analysis. Despite repeated attempts, only eight of these genes could be cloned in our vector system. Therefore, we focused our experimental analysis on this set of putative outer membrane proteins.

Five of them have been found in the outer membrane, confirming the ‘Hunter’ (Casadio et al., 2003) prediction, while just one was wrongly predicted; for one of this we got ambiguous results, and we could not determine if it is located in the outer membrane, finally we found that one of this could be an autotransporter.

References

Casadio R., Fariselli P., Finocchiaro G., Martelli P.L. (2003). Fishing new proteins in the twilight zone of genomes. The test case of outer membrane proteins in Escherichia coli K12, Escherichia coli O157:H7, and other Gram-negative bacteria, *Protein Sci* **12**:1158-1168.

Dowhan W. (1997) Molecular basis for membrane phospholipid diversity: why are there so many lipids?, *Annu Rev Biochem*. **66**: 199-232.

Henderson R., Unwin P.N.T., (1975) Three-dimensional model of purple membrane obtained by electron microscopy, *Nature* **257**: 28-32.

Madigan M.T., Martinko J.M., Parker J., Brock T.D. (VII Edition). Biology of Microorganisms (*Prentice Hall International*)

Schulz G.E. (2002) The structure of bacterial outer membrane proteins, *Biochim Biophys Acta* **1565** (2): 308-17.

Singer S.J. and Nicolson G.L. (1972) The fluid mosaic model of the structure of cell membranes, *Science,* **175** (23):720-730

von Heijne G. (1989) Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. *Nature* **341** (6241): 456-458

White S.H. and Wimley W.C. (1999) Membrane protein folding and stability: physical principles, *Annu Rev Biophys Biomol Struct,* 28:319-365

PROTEIN-DNA INTERACTIONS

CHAPTER 4

4.1 Biological Macromolecules: Protein and Acid nucleic

Two of the most important biological macromolecules are proteins and nucleic acids. The word *protein* comes from the Greek "*prota*", meaning "of primary importance". Proteins are large macromolecule composed of one or more polypeptide chains and are made of amino acids. All amino acids possess common structural features, an alpha carbon (C^α) that is connected to an amino (NH_3) group, a carboxyl group (COOH), and a variable side group (R) which gives each amino acids its distinctive chemical properties and helps to dictate the folding of the protein and are therefore critical to protein function (Figure 4).

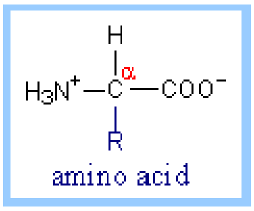


Figure 4
http://www.rpi.edu/dept/bcbp/molbiochem/MBWeb/mb1/part2/protein.htm

The amino acids in a polypeptide chains are covalently linked by peptide bonds formed in a dehydration reaction. Once linked in the protein chain, an individual amino acid is called a residue, and the linked series of carbon, nitrogen, and oxygen atoms are known as the main chain or protein backbone (http://en.wikipedia.org/wiki/Main_Page).

Proteins structure can be discussed in terms of four levels of complexity:

- The sequence of covalently linked amino acids is known as the primary structure of a protein.
- The secondary structure of a segment of polypeptide chain is the local spatial arrangement of its main-chain atoms without regard to the conformation of its side chains or its relationship with other segments. There are three common secondary structures in proteins: alpha helices, beta sheets and turns.
- Tertiary structure refers to the complete three dimensional folding of a protein and is maintained by more distant interactions.
- Quaternary structure is maintained by interchain interactions.

A nucleic acid carry genetic information or form structures within cells. The nucleic acids are very large molecules that have two main parts: the backbone that is made of alternating sugar and phosphate molecules bonded together in a long chain and nucleotide bases that are attached to the sugar groups in the backbone. Only four different nucleotide bases can occur in a nucleic acid, and the order in which they appear in the molecule is the coding for the information carried in the nucleic acid. The most common nucleic acids are deoxyribonucleic acid (DNA) and ribonucleic acid (RNA).

The four different nucleotide bases that occur in DNA are:
adenine (A), cytosine (C), guanine (G) and thymine (T); these bases are classified in to types: purines (adenine and guanine) and pyrimidines (cytosine and thymine).

The DNA comes in a double, complementary strand and the nucleotide base of on strand interacts with the other DNA strand in the helix in a specific way: adenine always bonds to thymine (and vice versa) with two hydrogen bonds and guanine always bonds to cytosine (and vice versa) with three hydrogen bonds (Figure 5).

The double helix is a right-handed spiral, and we can recognized a main and a minor groove. The narrowness of the minor groove means that the edges of the bases are more accessible in the major groove.

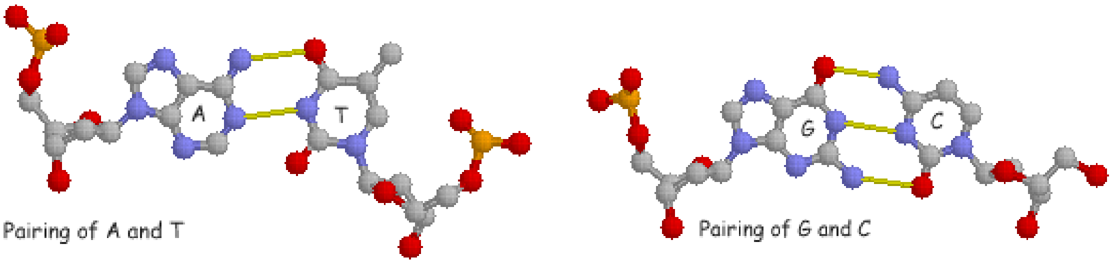


Figure 5

Base pairing. The hydrogen bonds between the NH (blue) and O (red) are in yellow.

http://www.rothamsted.ac.uk/notebook/courses/guide/dnast.htm

4.2 Database: PDB

The Protein Data Bank (PDB, Ref.1) has been founded by Edgar Meyer and Walter Hamilton from the Brookhaven National laboratory, then in 1998, the Research Collaboratory for Structural Bioinformatics (RCSB) has became responsible for the management of the PDB.

The PDB is the single worldwide repository for 3-D structures known with atomic resolution of large biological molecules, including proteins and nucleic acids.

These structures are obtained by X-ray crystallography or NMR spectroscopy

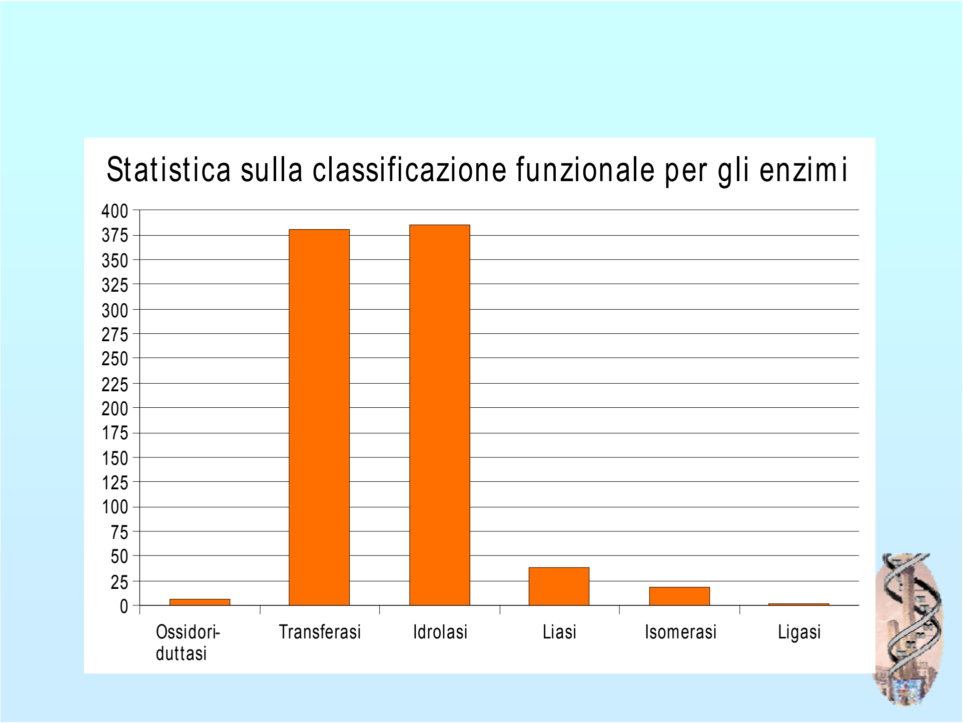
4.3 DNA-binding proteins

The most common and best studied DNA-binding proteins are the Zinc finger proteins, the Helix-turn-helix proteins, and the Leucine zipper proteins.

http://www.mun.ca/biochem/courses/3107/Topics/DNA_binding_proteins.html

4.4 Enzyme classification: EC number

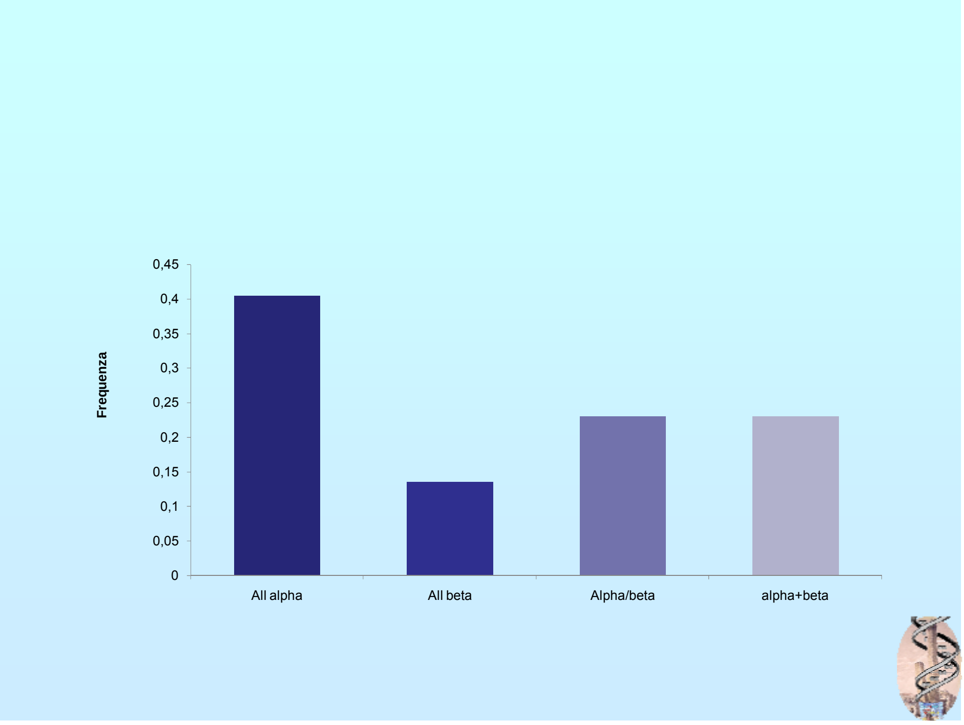
The Enzyme Commission number (EC number, Ref. 2) is a classification based on the chemical reactions that the enzyme catalyze.



Statistica sulla classificazione funzionale per gli enzimi

4.5 Structural classification of proteins

The Structural Classification of Proteins (SCOP, Ref. 3) database is a comprehensive ordering of all proteins of known structure, according to their evolutionary and structural relationships. The SCOP hierarchy comprises the following levels: Species, Protein, Family, Superfamily, Fold and Class.
This web site offer hierarchical classifications of the entire Protein Data Bank according to the folding patterns of the proteins.



References

Ref. 1 Protein Data Bank (http://www.rcsb.org/pdb/home/home.do)
Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E., (2000) The Protein Data Bank, *Nucleic Acids Research*, **28**: 235-242

Ref. 2 Enzyme nomenclature (http://www.chem.qmul.ac.uk/iubmb/enzyme/)

Ref. 3 Structural Classification of Proteins (http://scop.mrc-lmb.cam.ac.uk/scop/)

List of pubblications

- Marani P, Wagner S, Baars L, Genevaux P, de Gier JW, Nilsson IM, Casadio R, von Heijne G -New *Escherichia coli* Outer Membrane Proteins Identified Through Prediction And Experimental Verification- **Protein Science** 15:884-889 (2006)

- Casadio R, Calabrese R, Capriotti E, Compiani M, Fariselli P, Marani P, Montanucci L, Martelli PL, Rossi I, Tasco G -Machine learning and the prediction of protein structure: the state of the art- 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2004)- Perugia, 4-9/7/2004, Casa Editrice La Sapienza, Roma:933-940

- Casadio R, Calabrese R, Tasco G, Capriotti E, Compiani M, Marani P, Montanucci L, Rossi I, Martelli PL, Fariselli P-Metodi di Machine Learning per la predizione di strutture proteiche e della loro interazione- Convegno Bioinformatica: sfide e prospettive. Università del Sannio, 17-18/12/2003. F.Angeli Editore

Schools attended

- Bologna Winter School 2007: Bioinformatics For Systems & Synthetic Biology *New approaches and algorithms to cope with the future of Biological Science*, Bologna 18-23/02/2007
- Bologna Winter School 2006: Applied Bioinformatics: The Test Case Of Human Genome, Bologna 13-17/02/2006
- Scuola Nazionale di Biofisica 2005: Superfici E Biosistemi (XIII ciclo), Bressanone (BZ) 7-9/09/2005
- Bologna Winter School 2005: How Complex Is Functional Genomics?, Bologna, 13-19/02/2005
- Scuola Nazionale di Biofisica 2004: Proteomica e Biofisica (XII ciclo), Bressanone (BZ) 13-15/09/2004
- Bologna Winter School 2004**:** The State Of The Art Of Protein-Protein Interaction Networks: *The role of the in silico approach*, Bologna 8-14/02/2004
- Bologna Winter School 2003**:** Hot Topics In Structural Genomics', Bologna 9-15/02/2003

Attività didattica: tutor

Università di Bologna, Facoltà di Scienze MM.FF.NN.

anno accademico 2006/2007

Attività di tutorato per il “Laboratorio di Bioinformatica” Corso di Laurea in Biotecnologie.

Attività di tutorato per il “Banche dati medico-biologiche” Laurea Specialistica in Bioinformatica.

anno accademico 2005/2006

Attività di tutorato per il “Laboratorio di Biologia Computazionale 2” Corso di Laurea in Biotecnologie.

anno accademico 2004/2005

Attività di tutorato per il “Laboratorio di Biologia Computazionale 2” Corso di Laurea in Biotecnologie.

International Schools: local committee

2004- 2007: I have been part of the Local Committee at the International “Bologna Winter School” hold in Bologna.