



Biological Systems Workbook

Data modelling and simulations at molecular level

Edited by Javier Klett

First edition 2021

Javier Klett^{1,2}: The Structure of Organic Molecules, Structural Biology and Molecular Modelling sections.
ORCID ID: <https://orcid.org/0000-0001-8495-6568>

Carlos León¹: Databases in Molecular Biology section. ORCID ID: <https://orcid.org/0000-0001-8302-0995>

Bruno Di Geronimo²: revision of the text. ORCID ID: <https://orcid.org/0000-0003-1822-7142>

¹University Carlos III of Madrid. Department of Bioengineering and Aerospace Engineering, Madrid, Spain

²Experimental Therapeutics Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.

Book cover by Alfonso Núñez-Salgado

ISBN: 978-84-16829-65-1

Publisher: Carlos III University of Madrid

Av. Universidad 30, 28911 Leganés (Madrid) Spain

Electronic version available at the UC3M e-Archivo

<http://hdl.handle.net/10016/32421>

**This work is licensed under a
Creative Commons Attribution-NonCommercial-NoDerivatives 3.0 International License.**



uc3m | **Universidad Carlos III de Madrid**

This workbook is dedicated
to the family we are building,
to the family where I was born,
to the family I had chosen,
to the co-workers with whom I have coincided,
to those students who have endured me
and the rest who will do.

Index

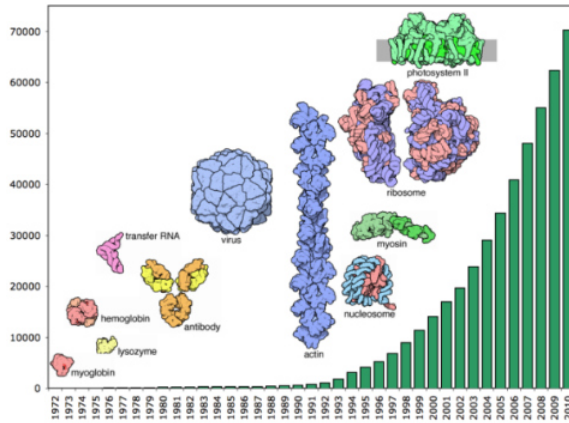
Introduction to Biological Systems.....	1
Biology and biomedicine need modelling.....	2
Workbook structure	3
The Structure of Organic Molecules.....	4
Molecular geometry	5
Atoms found in nature	5
Basic atoms in biomolecules (CHNOPS).....	5
Electron pair configuration and molecular geometry	6
Geometry of carbon	7
Geometry of nitrogen	8
Geometry of oxygen	8
Other atoms.....	9
Basic functional groups.....	9
Appendix functional groups	10
The SMILES notation	13
Examples	15
Molecular fingerprints	16
Examples.....	17
Acetic acid	17
Aspirin.....	17
Guanosine triphosphate (GTP).....	18
Molecular Force Fields.....	19
Atomic detail computer simulation.....	19
The potential energy of molecules	19
Molecular energy landscape.....	25
Entropic contribution.....	26
Molecular energy minimization	26
Minimization methods	27
Examples	28
Molecular similarity	29
Distances and similarity measures for small molecules	29
2D similarity, Tanimoto coefficient	30
Root-mean-square deviation (RMSD) of atomic positions.....	30
cRMSD: Structural alignment.....	31
Measuring similarity with ElectroShape.....	32
Conformational analysis	34
Pharmacophore modelling.....	36

Molecular space	36
Structural Biology	37
The role of sugars, lipids and other organic molecules.....	38
Carbohydrates	38
Lipids and phospholipids	42
Nucleic Acids, DNA and RNA	48
DNA.....	48
DNA packaging: nucleosomes and chromatin	49
Example: nucleosome	50
The genome.....	50
The genes.....	51
Genetics evolution.....	51
RNA	52
Proteins	53
Amino acids	53
Peptide units	54
The structure of proteins.....	55
Intrinsically disordered proteins.....	56
Protein complexes	58
Ramachandran plot.....	59
Storing the information of proteins	59
Experimental structures of macromolecules.....	60
X-ray crystallography	61
Electron microscopy	66
Protein NMR.....	68
Molecular Modelling.....	70
Molecular interactions	71
Bonded interactions	71
Non-bonded interactions	72
Binding between molecules.....	75
Molecular docking	78
Protein-protein docking.....	79
Protein-ligand docking.....	80
Drug discovery and rational drug design	82
Evaluating screening protocols	84
Protein similarity	84
Sequence alignment	85
Structural alignment of proteins	87

Protein contact map	87
Contact Overlap:	88
Structural classification of proteins	89
Homology modelling.....	90
Exploring the Motions of Biomolecules.....	91
Force fields and molecular mechanics	91
Molecular dynamics.....	93
Normal mode analysis and elastic network models	94
Databases in Molecular Biology.....	99
Introduction to databases	100
Introduction	100
What are (biological) databases?.....	101
What is (biological) data?.....	102
Characteristics and types of biological data	103
Why databases are essential in modern molecular biology	104
Potential pitfalls in databases	106
Nucleotides databases.....	107
Ensembl nucleotide database	107
Tools in nucleotide databases.....	110
Protein databases.....	112
Uniprot	112
Protein-protein interaction databases	115
Metabolite databases.....	117
Human metabolome database (HMDB)	117
References	119

INTRODUCTION TO BIOLOGICAL SYSTEMS

BIOLOGY AND BIOMEDICINE NEED MODELLING



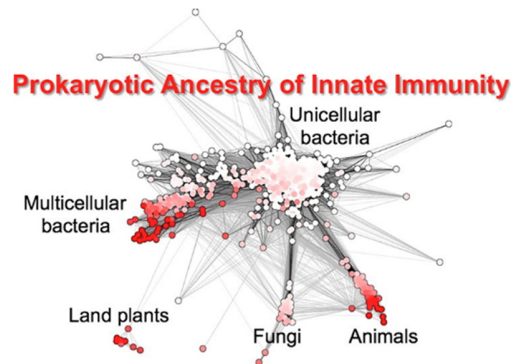
Time increase (x-axis) of the number of structures (y-axis) in the RCSB Protein Data Bank [1]. Image from the RCSB PDB (www.rcsb.org).

Hearth's ecosystem, such as the environmental perturbation related with the global warming, the increment of certain human or animal diseases, or the harassment and destabilization of ecosystems.

For this purpose, Mathematical and physical models help to find regularities in biological phenomena and to discover new biological laws, either qualitative explanations or quantitative predictions. Together with this, we need computational methods to calculate, store, compare, classify, and extrapolate all this data to other contexts and extract their whole potential.

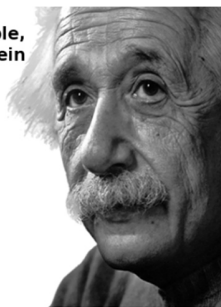
Nowadays, there are huge quantities of data surrounding the different fields of biology produced by high throughput experiments and theoretical simulations among other techniques. The results are often stored in biological data bases that are growing at a vertiginous rate every year [2]. Therefore, there is an increasing research interest in the application of mathematical and physical models to these Big Data problems. They are able to produce reliable predictions and explanations within its field of research.

For example, Next Generation Sequencing [4] is able to elucidate a bunch of genomic sequences in reduced time scales. High Throughput Screening [5] allows performing thousand biological assays in one row, etc. All this data is helping to overcome some biological questions and should push forward in the solution of problems faced by humans and the whole



Model of protein structure evolution [3]. (CC BY-NC-ND 3.0).

"Everything should be made as simple as possible, but not simpler" A Einstein



Ideographic representation of the different levels of complexity that could be applied while modelling a biological problem.

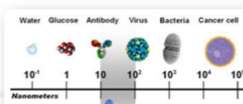
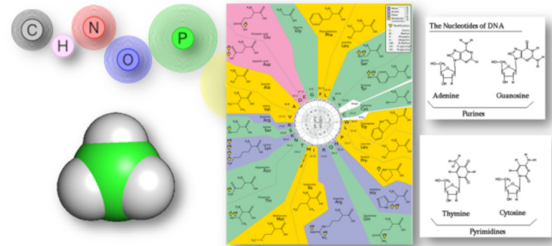
Simple qualitative models are very useful, without simplifications. The success of physics in simplifying its subjects of study shows that detailed predictions can be achieved only after a good qualitative understanding. Then, we can think that biological systems are like matryoshka dolls, where we have to choose the right level of description, avoiding both extreme reductionism and holism.

WORKBOOK STRUCTURE

In this Biological Systems Workbook, we aim to introduce the basic pieces allowing life to take place, from the 3D structural point of view. We will start learning how to look at the 3D structure of molecules from studying small organic molecules used as drugs. Meanwhile, we will learn some methods that help us to generate models of these structures. Then we will move to more complex natural organic molecules as lipid or carbohydrates, learning how to estimate and reproduce their dynamics. Later, we will revise the structure of more complex macromolecules as proteins or DNA. Along this process, we will refer to different computational tools and databases that will help us to search, analyze and model the different molecular systems studied in this course.

The structure of molecules

- Small molecules
- Basic biomolecules



The Structure of the Protein

Targeting Topoisomerases
 Interleukin drugs are used for cancer chemotherapy because their action cells that are rapidly dividing, like the cells in a growing tumor. Study of these drugs has revealed that **topoisomerases** are the major site of action. Topoisomerases begin by breaking DNA, then they make a topological change such as relieving supercoils or untangling strands, and finally they reconnect the DNA in its proper form. Interleukin drug block the reconnection step, leaving the topoisomerase after it has broken the DNA. This is a disaster when the cell divides: when a replication fork reaches the site, a lethal double-strand break is formed as the replication machinery passes through the damaged DNA.

Example of molecular system: Actinomycin

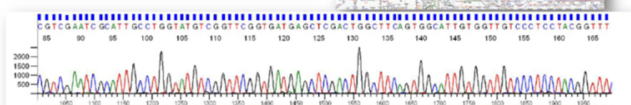
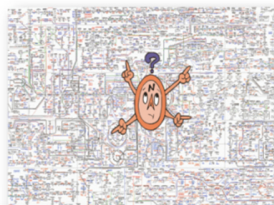
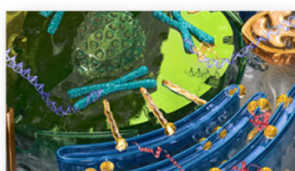
April 2003 lecture of the course by David Gooder.
<https://www.youtube.com/watch?v=20030401>
 Actinomycin D (A23187) is a potent anticancer drug. Data are water chemistry, and many times the search for medical compounds begins by looking to nature. Many antibiotics have been found by studying the complex, waxy structures between bacteria, and fungi, and isolating the toxic molecules that they build to protect themselves. Actinomycin is the first natural antibiotic discovered that has anti-cancer activity. It was discovered in the late 1940s by Japanese antibiotic chemists. Unfortunately, it is too toxic for general use. Using cancer cells that also possess the genes, but related molecules have subsequently been discovered, and are now widely used for cancer chemotherapy.

Structural biology

- DNA
 - Representing the DNA structure
 - DNA as a medical target
- Proteins
 - Representing the structure of proteins
- The role of sugars, lipids and other organic molecules.

Molecular Modelling

- Docking
 - Molecular interactions
 - Molecular association models
 - Docking proteins and small molecules
- Exploring the motions of molecules
 - MD
 - NMA
- Modelling the 3D structure of proteins
 - Comparing the structure of proteins
 - Ab-initio modelling
 - Homology modelling
 - Structural classification of proteins



Data Bases in molecular biology

- Review of molecular biology concepts
- Methods for molecular biology massive-data generation
- Gene, protein and metabolite databases and tools

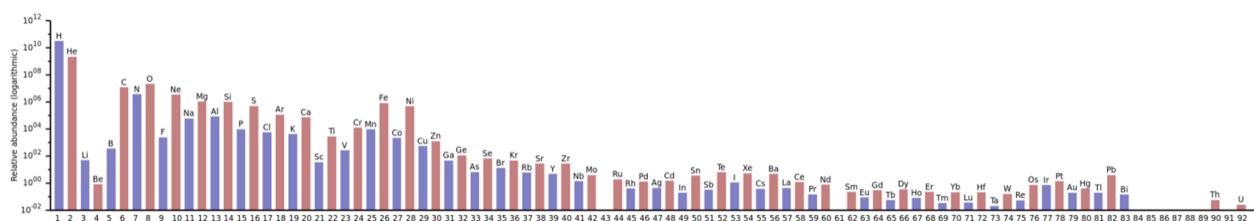
THE STRUCTURE OF ORGANIC MOLECULES

MOLECULAR GEOMETRY

ATOMS FOUND IN NATURE

In the periodic table, there are 118 atoms defined, however, from those, only 92 are found in nature. The rest of them have been artificially created in laboratories. Hydrogen and helium atoms are commonly found in the universe. The nuclear fusion of hydrogen into helium powers the majority of stars from the entire cosmos since the Big Bang. From the 92 elements found in nature, all of them were formed in generations of stars.

During the life of a star, its fuel is burned into heavier elements. After the star dies, elements are released as enriched elements back into the cosmos. When the heavier elements become abundant enough, they can form rocky planets as the planet Earth among other astronomical objects. Then, it is common to detect those 92 atom types in cosmic objects, but they decrease in the presence of elements as they become heavier, with some exceptions as Iron, that is common since it is the minimum energy nuclide made by fusion of helium in supernovae [6]. This is known as the Oddo-Harkins rule that states that an element with an odd atomic number, i.e. with an unpaired proton, is likely to capture another proton.

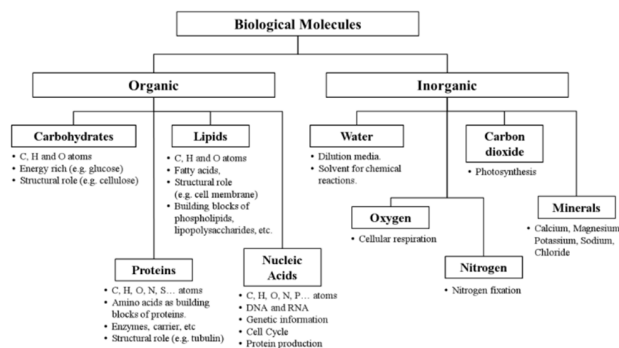
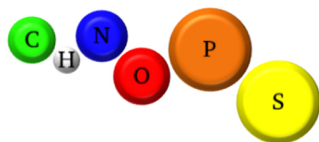


Estimated abundances of the chemical elements in the Solar system. Image from Wikimedia Commons [7](CC BY-NC-ND 3.0).

BASIC ATOMS IN BIOMOLECULES (CHNOPS)

CHNOPS is an acronym used for common atoms found in organic molecules: Carbon, Hydrogen, Nitrogen, Oxygen, Phosphorus and Sulfur. They differ in size and in the number of double and single bonds they can form, i.e. atomic geometry when forming a compound.

For example, Carbon, which is the most common atom, it can be bonded to four different atoms through single bonds, or to 3 atoms by one double and two single bonds, or even to two atoms by one triple and one double bond or two double bonds. From these different bonding combinations within the different atom types, a great variety of geometries and shapes can arise, describing the structure of the molecules. The more common geometries include linear, planar, or forming a tetrahedron. Most of the different organic molecules found in nature are built out of these simple atom types and these simple geometries. For example, only carbon, hydrogen, some oxygen and nitrogen atoms usually form simple carbohydrates and lipids.



Left, the relative atomic diameters of carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur represented in CPK colors [7]. Right, schema with examples of some organic and inorganic molecules.

CPK coloring [8] is a convention for distinguishing atoms of different chemical elements in molecular models designed by chemists Robert Corey and Linus Pauling, and further developed by Walter Koltun. The CPK color code defines carbon in black, nitrogen in blue, oxygen in red, phosphorus in orange, sulfur in yellow, etc. Note that for representation purposes, carbon is also used as a “wildcard” having any color different from red or blue.

In the following sections, we will revise the geometry of carbon, oxygen, nitrogen and phosphorus and sulfur atoms, as they are the main building blocks of biomolecules that we are going to study along this course.

Types of bonding arrangements			
Element	~ Radius (Å)	All single bonds	Other Combinations
C	0.77	4	2 single + 1 double 2 double 1 single + 1 triple
H	0.37	1	
N	0.70	3 Positively charged 4	1 double + 2 singles, Positively charged: 2 double + 2 single 1 triple + 1 single
O	0.66	2	1 double Negatively charged: 1 single
P	1.10	3, 5, Negatively charged 4...	1 double + 3 single...
S	1.04	2, positively charged 3	2 double and 2 single

Number of atom, bonds and bond types of Carbon, Nitrogen, Oxygen and Hydrogen atoms

ELECTRON PAIR CONFIGURATION AND MOLECULAR GEOMETRY

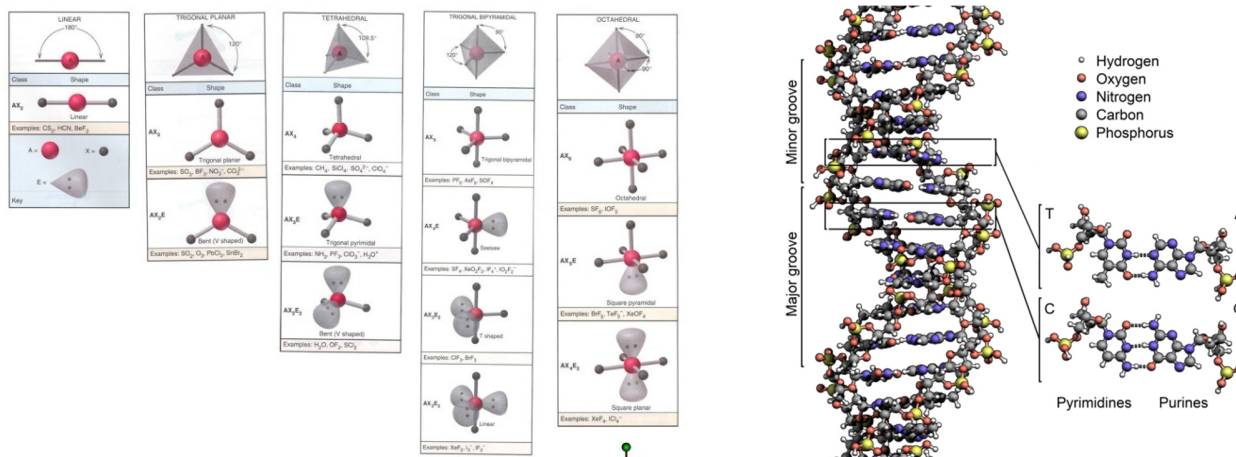
Along this course, we are going to focus on the structure of biomolecules. They will adopt complex shapes formed, in many cases, by a repetition of some geometrical elements. These geometrical elements are derived from the electron pair configuration of each single atom. Then, if we want to understand the complexity of molecules as proteins or DNA, it is necessary to pay attention on how its atoms are bound and located in the molecule.

Apart from the nature of the bonds between atoms, single/double/triple bond, when atoms are bound between them, they can adopt linear, planar, tetrahedral, bipyramidal and octahedral geometries, depending on the electron pair configuration of the atoms implied in the bond.

Other important aspect to take into account is that only single bonds will allow rotations giving the flexibility to the molecule, i.e., the degrees of freedom of a molecule can be resumed to the number of rotatable bonds found in the molecule.

For example, DNA is formed by nucleotides, composed by hydrogen, carbon, oxygen, nitrogen and phosphorus atoms. Then, nucleotides are bound between them by the phosphodiester linkage, that rotates some these specific bonds allowing the staking and hydrogen bonding interactions between the base pairs. All this, results in the characteristic helical configuration of DNA.

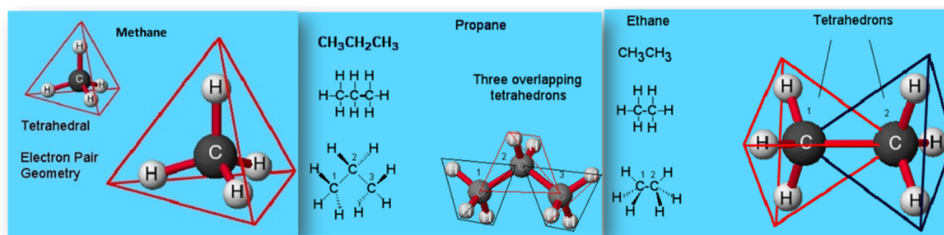
In the following sections, we are going to revise the 3D geometry of carbon, oxygen, nitrogen and phosphorus and sulfur atoms, as they are the main building blocks of biomolecules that we are going to study along this course



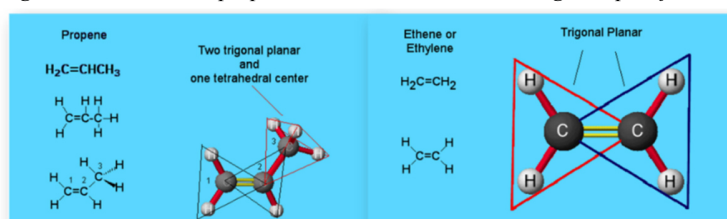
Left, list of the electron pair configuration observed in atoms belonging to biomolecules. Image from Virtual Chembook [9]. Right, representation of the three dimensional structure, of a DNA molecule. Image from Wikimedia Commons [7] (CC BY-NC-ND 3.0)..

GEOMETRY OF CARBON

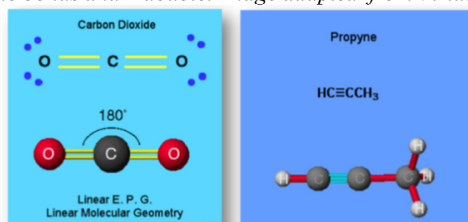
Carbon is one of the most abundant atoms on earth and the main pillar of the organic chemistry. It can adopt tetrahedral, planar or linear geometries depending on the number of double or single bonds. A carbon atom will have to share 4 free electrons with other atoms as shown in the figures below.



Tetrahedral, "4 single bonds" methane, propane and ethane molecules. Image adapted from Virtual Chembook [9].



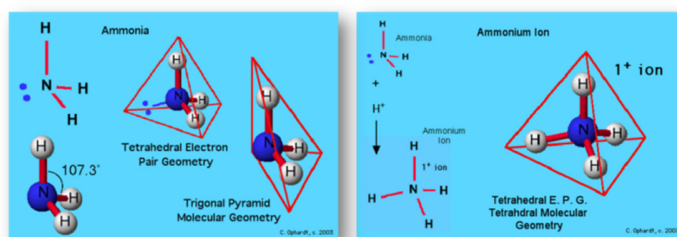
Planar, 2 single bonds and 1 double. Image adapted from Virtual Chembook.[9].



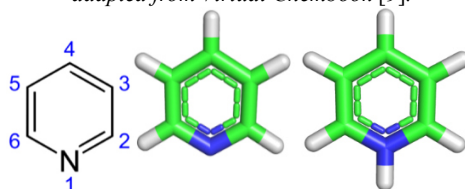
Linear "2 double bonds" and "1 single and 1 triple bond". Image adapted from Virtual Chembook [9].

GEOMETRY OF NITROGEN

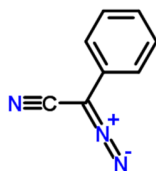
Nitrogen can bind to almost any atom in the periodic table and it is one of most important atoms in organic chemistry. It can also adopt tetrahedral, planar or linear geometries depending on the number of double or hydrogen bonds. Also, Nitrogen can be neutral or protonated (positively charged), allowing it to have up to four single bonds or 1 double and 2 single when protonated. In both situations, it has a tetrahedral configuration as it belongs to an aromatic ring, conferring it a planar geometry due to the circulation of the pi-electrons.



Tetrahedral. Left, neutral “3 single bond” Nitrogen atom. Right, protonated/”positively charged” “4 single bonds” Nitrogen atom. Image adapted from Virtual Chembook [9].



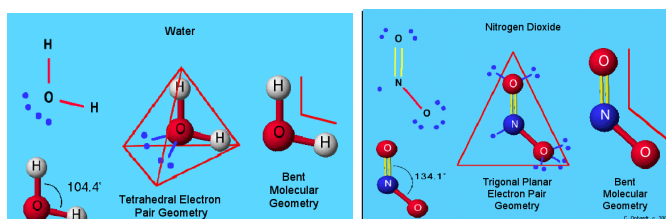
Planar. Left, two dimensional representation of a pyrimidine with implicit hydrogen atoms representation. Middle, neutral “1 single and 1 double bond” pyridine. Right, “positively charged” “2 single 1” double bond pyridine.



Linear Left, Diazo(phenyl)acetonitrile showing “1 triple bond” and “2 double bonds” in a neutral and protonated/”positively charged” Nitrogen atoms respectively

GEOMETRY OF OXYGEN

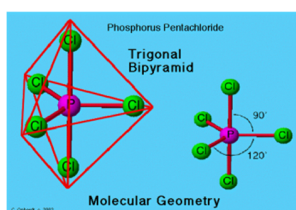
Oxygen is a reactive atom and an oxidizing agent that easily forms oxides with other molecules surrounding it. When neutral or negatively charged, i.e. with two or only one single bond, it will adopt a tetrahedral geometry. However, when having only one double bond it will have a planar geometry because the two remaining electron pairs.



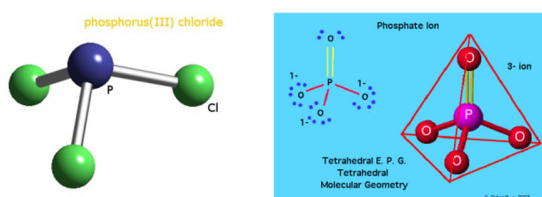
Left, Tetrahedral/bent “2 single bonds” Oxygen atom of a water molecule. Right a nitrogen dioxide molecule with two oxygen atoms, the top one with one single bond planar and the other tetrahedral and negatively charged with only one single bond. Image adapted from Virtual Chembook [9].

OTHER ATOMS

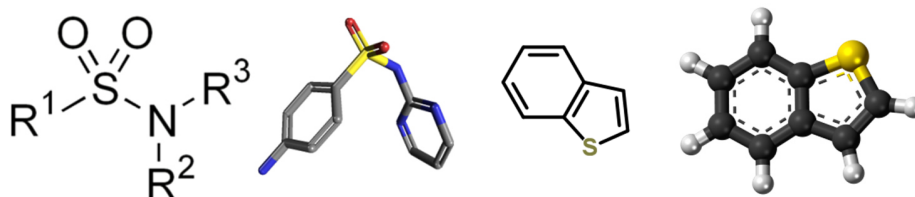
There are other atoms that are less frequent but very important for biological process as Phosphorus or Sulfur that appear in molecules as GTP, DNA, amino acids, etc.



Trigonal Bipyramid, phosphorus atom with “5 single bond”. Image adapted from Virtual Chembook [9].



Tetrahedral, phosphorus atom with “3 single bond” and “3 single and 1 double bond”. Image adapted from Virtual Chembook [9].



Left, tetrahedral “2 double 2 single bonds” sulfur atom in a Sulfadiazine. Right, “2 single bond sulfur” atom in a Benzothiophene.

BASIC FUNCTIONAL GROUPS

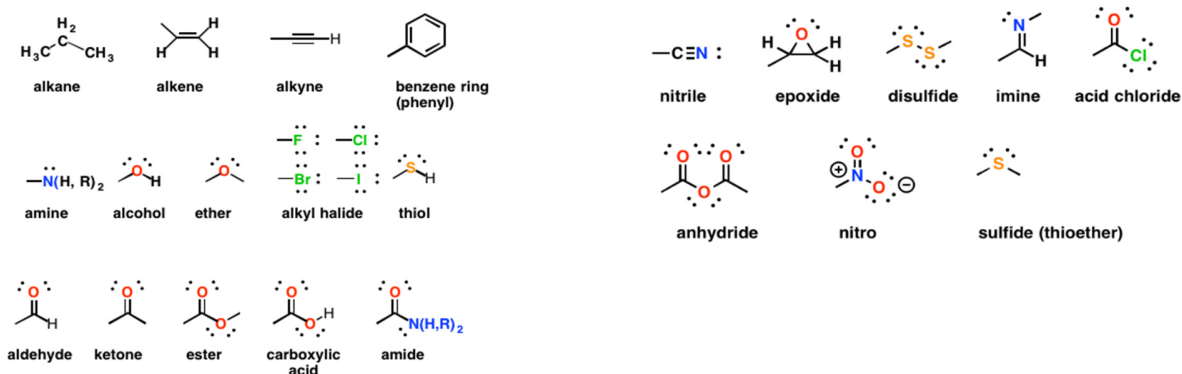
In organic chemistry, there are certain sets of atoms that are usually found over the molecular space. They typically have a certain chemical property that let us refer to them with a certain group name, helping to the understanding and the description of its chemistry. Then, we could define functional groups as specific sets of atoms and bonds within molecules that are responsible for the characteristic of those molecules. Normally these functional groups have reaction chemistry independent on the environment.

Functional groups are considered as groups of atoms within molecules with characteristic properties. There are different combinations, but they all help us to refer to these common fragments or regions of molecules, to understand the interactions that they may form and their reactivity among other properties. This approach is empirically determined by experience and experimental results but does not rely in complex theoretical approaches as Valence shell electron pair repulsion (VSEPR)[10] theory or the molecular orbital (MO) theory [11].



Left, Benzyl acetate with an ester functional group (in red), an acetyl mo, and a benzyloxy moiety (circled with light orange). Image from Wikimedia Commons [7](CC BY-NC-ND 3.0). Right, morphine with several functional groups and chiral. Image from Mark R. Leach [12].

Some common examples of functional groups comprehend alcohols, amines, carboxylic acids, ketones, and ethers. Depending on the level of specificity of the definition, there may be even hundreds of them, but only 14 of them may be found in most of organic molecules. Then we could extend that 14 to another 8 that are also common in the molecular space.



Left, list of 14 functional groups considered as key players. Right, 8 extra functional groups also commonly observed. Image adapted from "Organic Chemistry with a Biological Emphasis (Soderberg)" [13] (CC BY-NC-ND 3.0).

APPENDIX FUNCTIONAL GROUPS

To be proficient in organic chemistry at university entrance level exam systems (i.e., American AP, British A-Level or French Baccalaureate), students should be able to recognize the 30 functional group listed below [12].

Alkane

R

Alkyl functions are represented by the R-
Methyl: CH_3 ?
Ethyl: CH_3CH_2 ?
Propyl: $\text{CH}_3\text{CH}_2\text{CH}_2$?
Isopropyl: $(\text{CH}_3)_2\text{CH}$?
Phenyl: C_6H_5 ?
etc.

Primary alcohol



Primary alcohols have an -OH function attached to an R-CH₂- group.
Primary alcohols can be oxidised to aldehydes and on to carboxylic acids. (It can be difficult to stop the oxidation at the aldehyde stage.)
Primary alcohols can be shown in text as: RCH₂OH

Secondary alcohol



Secondary alcohols have an -OH function attached to a R₂CH- group.
Secondary alcohols can be oxidised to ketones.
Secondary alcohols can be shown in text as: R₂CHOH

Tertiary alcohol



Tertiary alcohols have an -OH function attached to a R₃C- group.
Tertiary alcohols are resistant to oxidation with acidified potassium dichromate(VI), K.
Tertiary alcohols can be shown in text as: R₃COH

Carbonyl function



The carbonyl group is a super function because many common functional groups are based on a carbonyl attached to two different radicals, including:
aldehydes, ketones, carboxylic acids, esters, amides, acyl (acid) chlorides, acid anhydrides

Aldehyde



Aldehydes have a hydrogen and an alkyl (or aromatic) group attached to a carbonyl function.

Aldehydes can be shown in text as: RCHO

Aldehydes are easily oxidised to carboxylic acids, and they can be reduced to primary alcohols.

Aldehydes can be distinguished from ketones by giving positive test results with Fehlings solution (brick red precipitate) or Tollens reagent (silver mirror).

Aldehydes give red-orange precipitates with 2,4-dinitrophenyl hydrazine.

Ketone



Ketones are carbonyl groups attached to a pair of alkyl or aromatic groups.

Ketones can be shown in text as: RCOR

Ketones can be distinguished from aldehydes by giving negative test results with Fehling's solution (brick red precipitate) or Tollens reagent (silver mirror).

Ketones give red-orange precipitates with 2,4-dinitrophenyl hydrazine.

Carboxylic acid



Carboxylic acids have an alkyl or aromatic groups attached to a hydroxy-carbonyl function.

Carboxylic acids can be shown in text as: RCOOH

Carboxylic acids are weak Bronsted acids, they liberate CO₂ from carbonates, and hydrogen carbonates.

Ester



Esters have a pair of alkyl or aromatic groups attached to a carbonyl + linking oxygen function.

Esters can be shown in text as: RCOOR or (occasionally) ROCOR.

carboxylic acid + alcohol -> ester + water

This is an acid catalyzed equilibrium.

Amide



Primary amides (shown) have an alkyl or aromatic group attached to an amino-carbonyl function.

Primary amides can be shown in text as: RCONH₂

Secondary amides have an alkyl or aryl group attached to the nitrogen: RCONHR

Tertiary amides have two alkyl or aryl group attached to the nitrogen: RCONR₂

Primary amine



Primary amines have an alkyl or aromatic group and two hydrogens attached to a nitrogen atom.

Primary amines can be shown in text as: RNH₂

Primary amines are basic functions that can be protonated to the corresponding ammonium ion.

Primary amines are also nucleophilic.

Secondary amine



Secondary amines have a pair of alkyl or aromatic groups, and a hydrogen, attached to a nitrogen atom.

Secondary amines can be shown in text as: R₂NH

Secondary amines are basic functions that can be protonated to the corresponding ammonium ion. Secondary amines are also nucleophilic.

Tertiary amine



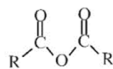
Tertiary amines have three alkyl or aromatic groups attached to a nitrogen atom.

Tertiary amines can be shown in text as: R₃N

Tertiary amines are basic functions that can be protonated to the corresponding ammonium ion.

Tertiary amines are also nucleophilic.

Acid anhydride



Acid anhydrides are formed when water is removed from a carboxylic acid, hence the name.

Acid anhydrides can be shown in text as: (RCO)₂O

Nitrile



Nitriles (or organo cyanides) have an alkyl (or aromatic) group attached to a carbon-triple-bond-nitrogen function.

Nitriles can be shown in text as: RCN

Note that there is a nomenclature issue with nitriles/cyanides. If a compound is named as the nitrile then the nitrile carbon is counted and included, but when the compound is named as the cyanide it is not.

For example: CH₃CH₂CN is called propane nitrile or ethyl cyanide (cyanoethane).

Carboxylate ion or salt



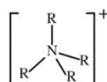
Carboxylate ions are the conjugate bases of carboxylic acids, i.e. the deprotonated carboxylic acid.

Carboxylate ions can be shown in text as: RCOO⁻

When the counter ion is included, the salt is being shown.

Salts can be shown in text as: RCOONa

Ammonium ion

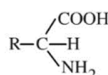


Ammonium ions have a total of four alkyl and/or hydrogen functions attached to a nitrogen atom. [NH₄]⁺

[RNH₃]⁺ [R₂NH₂]⁺ [R₃NH]⁺ [R₄N]⁺

Quaternary ammonium ions are not proton donors, but the others are weak Bronsted acids (pK_a about 10).

Amino acid

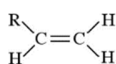


Amino acids, strictly α-amino acids, have carboxylic acid, amino function and a hydrogen attached to the same carbon atom.

There are 20 naturally occurring amino acids. All except glycine (R = H) are chiral and only the L enantiomer is found in nature.

Amino acids can be shown in text as: R-CH(NH₂)COOH

Alkene



Alkenes consist of a C=C double bond function.

Alkenes can be shown in text as:

Mono

substituted:

RCH=CH₂

1,1-disubstituted:

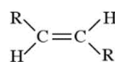
R₂C=CH₂

1,2-disubstituted: RCH=CHR

Alkenes are planar and there is no rotation about the C=C bond.

Alkenes are electron rich reactive centers and are susceptible to electrophilic addition.

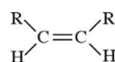
trans-Alkene



trans-alkenes are 1,2-disubstituted functions with the two R, X or other groups on opposite sides of the C=C function.

Due to the non-rotation of the C=C bond, cis and trans geometric isomers are not [thermally] interconverted.

cis-Alkene



cis-Alkenes are 1,2-disubstituted functions with the two R, X or other groups on the same side of the C=C function.

Due to the non-rotation of the C=C bond, cis and trans geometric isomers are not [thermally] interconverted.

Ether

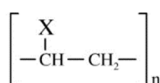


Ethers have a pair of alkyl or aromatic groups attached to a linking oxygen atom.

Ethers can be shown in text as: ROR

Ethers are surprisingly unreactive and are very useful as solvents for many many (but not all) classes of reaction.

Polymer



Polymers consist of small monomer molecules that have reacted together so as to form a large covalently bonded structure.

There are two general types of polymerization: addition and condensation.

Linear chain polymers are generally thermoplastic, while three dimensional network polymers are not.

Table adapted from Mark R. Leach [12].

THE SMILES NOTATION

Describing the structure of chemical in form of a single string is of particular interest in medicinal chemistry. There are different notations for that purpose, but the simplified molecular-input line-entry system (SMILES) is one of the most used. SMILES strings are used over most molecular structure editors and web servers for transforming 2D drawings into a 3D model molecule, and *vice versa*.

The initial SMILES specification was defined in 1980s, and it has been modified, improved and extended over the years. OpenSMILES is a project within the Blue Obelisk community [14] formed by a group of chemists, programmers, informaticians, etc.; devoted to open-source computational developments and standard definition in chemoinformatics. OpenSMILES defines a complete grammar summarized below.

Section	Formal Grammar
ATOMS	
<u>Atoms</u>	<code>atom ::= bracket_atom aliphatic_organic aromatic_organic '*'</code>
ORGANIC SUBSET ATOMS	
<u>Organic Subset</u>	<code>aliphatic_organic ::= 'B' 'C' 'N' 'O' 'S' 'P' 'F' 'Cl' 'Br' 'I'</code> <code>aromatic_organic ::= 'b' 'c' 'n' 'o' 's' 'p'</code>
BRACKET ATOMS	
<u>Bracket Atoms</u>	<code>bracket_atom ::= '[' isotope? symbol chiral? hcount? charge? class? ']'</code> <code>symbol ::= element_symbols aromatic_symbols '*'</code> <code>isotope ::= NUMBER</code> <code>element_symbols ::= 'H' 'He' 'Li' 'Be' 'B' 'C' 'N' 'O' 'F' 'Ne' 'Na' 'Mg' 'Al' 'Si' 'P' 'S' 'Cl' 'Ar' 'K' 'Ca' 'Sc' 'Ti' 'V' 'Cr' 'Mn' 'Fe' 'Co' 'Ni' 'Cu' 'Zn' 'Ga' 'Ge' 'As' 'Se' 'Br' 'Kr' 'Rb' 'Sr' 'Y' 'Zr' 'Nb' 'Mo' 'Tc' 'Ru' 'Rh' 'Pd' 'Ag' 'Cd' 'In' 'Sn' 'Sb' 'Te' 'I' 'Xe' 'Cs' 'Ba' 'Hf' 'Ta' 'W' 'Re' 'Os' 'Ir' 'Pt' 'Au' 'Hg' 'Tl' 'Pb' 'Bi' 'Po' 'At' 'Rn' 'Fr' 'Ra' 'Rf' 'Ob' 'Sg' 'Bh' 'Hs' 'Mt' 'Ds' 'Rg' 'Cn' 'Fl' 'Lv' 'La' 'Ce' 'Pr' 'Nd' 'Pm' 'Sm' 'Eu' 'Gd' 'Tb' 'Dy' 'Ho' 'Er' 'Tm' 'Yb' 'Lu' 'Ac' 'Th' 'Pa' 'U' 'Np' 'Pu' 'Am' 'Cm' 'Bk' 'Cf' 'Es' 'Fm' 'Md' 'No' 'Lr'</code> <code>aromatic_symbols ::= 'b' 'c' 'n' 'o' 'p' 's' 'se' 'as'</code>
CHIRALITY	
<u>Chirality</u>	<code>chiral ::= '@' '@@' '@TH1' '@TH2' '@AL1' '@AL2' '@SP1' '@SP2' '@SP3' '@TB1' '@TB2' '@TB3' ... '@TB20' '@OH1' '@OH2' '@OH3' ... '@OH30' '@TB DIGIT DIGIT' '@OH DIGIT DIGIT'</code>
HYDROGENS	
<u>Hydrogens</u>	<code>hcount ::= 'H' 'H' DIGIT</code>
CHARGES	
<u>Charge</u>	<code>charge ::= '-' '.' DIGIT? DIGIT '+' '*' DIGIT? DIGIT '-..' deprecated '+..' deprecated</code>
ATOM CLASS	
<u>Atom Class</u>	<code>class ::= ':' NUMBER</code>
BONDS AND CHAINS	
<u>Bonds</u>	<code>bond ::= '-' '=' '#' '\$' ':' '/' '\''</code> <code>ringbond ::= bond? DIGIT bond? '%' DIGIT DIGIT</code> <code>branched_atom ::= atom ringbond* branch*</code> <code>branch ::= '(' chain ')' '(' bond chain ')' '(' dot chain ')'</code> <code>chain ::= branched_atom chain branched_atom chain bond branched_atom chain dot branched_atom</code> <code>dot ::= '.'</code>
SMILES STRINGS	
	<code>smiles ::= terminator chain terminator</code>

An atom is represented by its atomic symbol, enclosed in square brackets, []. The first character of the symbol is uppercase and the second (if any) is lowercase, except that for aromatic atoms where the first character is lowercase. The symbol '*' is also accepted as a valid atomic symbol, and represents a "wild-card" or unknown atom.

SMILES	Atomic Symbol
[U]	Uranium
[Pb]	Lead
[He]	Helium
[*]	Unknown atom

Multiple hydrogens are represented inside of brackets as *Hn* where n is a number such as H3. Note that, [C] and [CH0] are identical, and [CH] and [CH1] are identical.

SMILES	Name	Comments
[CH4]	methane	
[ClH]	hydrochloric acid	H1 implied
[ClH1]	hydrochloric acid	

Charge is specified by a +n or -n where n is a number, if the number is missing, it means either +1 or -1 as appropriate.

SMILES	Name	Comments
[Cl-]	chloride anion	-1 charge, H0 implied
[OH1-]	hydroxyl anion	-1 charge, H1
[OH-1]	hydroxyl anion	-1 charge, H1
[Cu+2]	copper cation	+2 charge, H0 implied
[Cu++]	copper cation	+2 charge, H0 implied


The "organic subset" of B, C, N, O, P, S, F, Cl, Br, I, and * (the "wildcard" atom) can be used by only the atomic symbol. The "organic subset" notation assumes the following properties: 1) "implicit hydrogens" are added such as the valence of the atom is in the lowest, 2) the atom's charge is zero, 3) the atom has no isotopic specification, and 4) the atom has no chiral specification.

SMILES	Name
C	methane
N	ammonia
Cl	hydrochloric acid



Atoms that are adjacent in a SMILES string are assumed to be joined by a single or aromatic bond. Double, triple and quadruple bonds are represented by '=', '#', and '\$' respectively.

SMILES	Name	SMILES	Name
CC	ethane	C=C	ethene
CCO	ethanol	C#N	hydrogen cyanide
NCCCC	n-butylamine	CC#CC	2-butyne
CCCCN	n-butylamine	CCC=O	propanol
		[Rh-](Cl)(Cl)(Cl)(Cl)\$[Rh-](Cl)(Cl)(Cl)Cl	octachlorodirhenate (III)




An atom with three or more bonds is called a branched atom, and it is represented using parentheses.

Depiction	SMILES	Name
	CCC(CC)CO	2-ethyl-1-butanol

Branches can be nested or "stacked" to any depth.

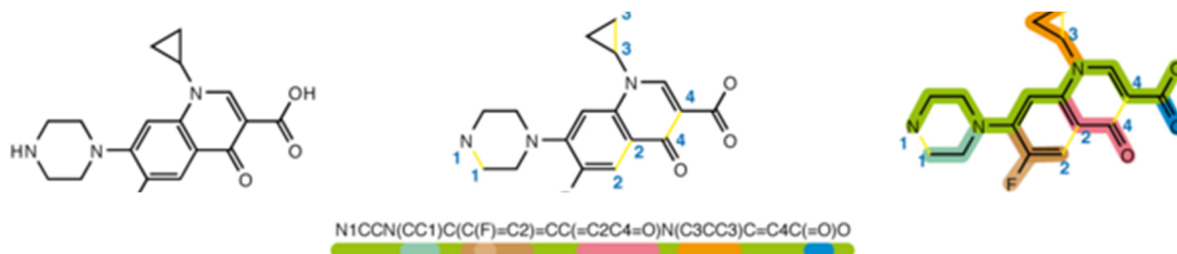
Depiction	SMILES	Name
	CC(C)C(=O)C(C)C	2,4-dimethyl-3-pentanone
pic here	OCC(CCC)C(C(C)C)CCC	2-propyl-3-isopropyl-1-propanol
	OS(=O)(=S)O	thiosulfate

In a string such as "C1CCCCC1", the first occurrence of a ring-closure number (an "rnum") creates an "open bond" to the atom that precedes the ring-closure number (the "rnum"). When that same rnum is encountered later in the string, a bond is made between the two atoms, that typically forms a cyclic structure.

Depiction	SMILES	Name
	C1CCCCC1	cyclohexane
	N1CC2CCCC2CC1	perhydroisoquinoline
	C=1CCCCC=1	cyclohexene
	C=1CCCCC1	cyclohexene (preferred form)
	C1CCCCC=1	cyclohexene
	C-1CCCCC=1	invalid

EXAMPLES

Ciprofloxacin, a fluoroquinolone antibiotic.



Deriving the SMILES representation of a chemical molecule. Image from Wikimedia Commons [7] (CC BY-NC-ND 3.0)

Other examples:

Molecule	Structure	SMILES Formula
Dinitrogen	$N \equiv N$	<chem>N#N</chem>
Methyl isocyanate (MIC)	$CH_3-N=C=O$	<chem>CN=C=O</chem>
Copper(II) sulfate	$Cu^{2+} SO_4^{2-}$	<chem>[Cu+2].[O-]S(=O)(=O)[O-]</chem>
Vanillin		<chem>O=Cc1cc(O)c(OC)c1</chem> <chem>OCc1cc(C=O)ccc1O</chem>
Melatonin ($C_{13}H_{16}N_2O_2$)		<chem>CC(=O)NCCC1=CNC2c1cc(OC)cc2</chem> <chem>CC(=O)NCCc1c[nH]c2ccc(OC)cc12</chem>

SMILES notation of Dinitrogen, Methyl isocyanate, Copper(II) sulfate, Vanillin and Melatonin molecules. Table from Wikimedia Commons [7] (CC BY-NC-ND 3.0).

MOLECULAR FINGERPRINTS

In chemoinformatics, the Molecular Fingerprint of a certain compound is a simple definition of the structure and chemical properties in terms of a bit string. While comparing the 2D or 3D structures is challenging, comparing and searching for bit strings is well defined and optimized in computer sciences. Then molecular fingerprints facilitate the integration of big-data set information into molecular databases, allowing their comparison and their substructure search.

A simple example of molecular fingerprints is, "Structural fingerprints" derived from sub-structural properties and "hash fingerprints" that are used to encode the 2D structural characteristics into a bit vector. In a structural fingerprint, each bit corresponds to the presence or count of individual chemical atoms, a chemical property or substructure with a certain chemical relevance [15].

<p>Section 1: Hierarchic Element Counts - These bits test for the presence or count of individual chemical atoms represented by their atomic symbol.</p> <table border="1"> <thead> <tr> <th>Position</th> <th>Bit-Substructure</th> </tr> </thead> <tbody> <tr><td>0</td><td>>= 4 H</td></tr> <tr><td>1</td><td>>= 8 H</td></tr> <tr><td>2</td><td>>= 16 H</td></tr> <tr><td>3</td><td>>= 32 H</td></tr> <tr><td>4</td><td>>= 1 Li</td></tr> <tr><td>5</td><td>>= 2 Li</td></tr> <tr><td>6</td><td>>= 1 B</td></tr> <tr><td>7</td><td>>= 2 B</td></tr> <tr><td>8</td><td>>= 4 B</td></tr> <tr><td>9</td><td>>= 2 C</td></tr> <tr><td>10</td><td>>= 4 C</td></tr> <tr><td>11</td><td>>= 8 C</td></tr> <tr><td>12</td><td>>= 16 C</td></tr> <tr><td>...</td><td></td></tr> </tbody> </table>	Position	Bit-Substructure	0	>= 4 H	1	>= 8 H	2	>= 16 H	3	>= 32 H	4	>= 1 Li	5	>= 2 Li	6	>= 1 B	7	>= 2 B	8	>= 4 B	9	>= 2 C	10	>= 4 C	11	>= 8 C	12	>= 16 C	...		<p>Section 2: Rings in a canonic Extended Smallest Set of Smallest Rings (ESSSR) ring set - These bits test for the presence or count of the described chemical ring system. An ESSSR ring is any ring which does not share three consecutive atoms with any other ring in the chemical structure. For example, naphthalene has three ESSSR rings (two phenyl fragments and the 10-membered envelope), while biphenyl will yield a count of only two ESSSR rings.</p> <table border="1"> <thead> <tr> <th>Bit Position</th> <th>Bit Substructure</th> </tr> </thead> <tbody> <tr><td>115</td><td>>= 1 any ring size 3</td></tr> <tr><td>116</td><td>>= 1 saturated or aromatic carbon-only ring size 3</td></tr> <tr><td>117</td><td>>= 1 saturated or aromatic nitrogen-containing ring size 3</td></tr> <tr><td>118</td><td>>= 1 saturated or aromatic heteroatom-containing ring size 3</td></tr> <tr><td>119</td><td>>= 1 unsaturated non-aromatic carbon-only ring size 3</td></tr> <tr><td>120</td><td>>= 1 unsaturated non-aromatic nitrogen-containing ring size 3</td></tr> <tr><td>121</td><td>>= 1 unsaturated non-aromatic heteroatom-containing ring size 3</td></tr> <tr><td>...</td><td></td></tr> <tr><td>129</td><td>>= 1 any ring size 4</td></tr> <tr><td>130</td><td>>= 1 saturated or aromatic carbon-only ring size 4</td></tr> <tr><td>131</td><td>>= 1 saturated or aromatic nitrogen-containing ring size 4</td></tr> <tr><td>132</td><td>>= 1 saturated or aromatic heteroatom-containing ring size 4</td></tr> <tr><td>133</td><td>>= 1 unsaturated non-aromatic carbon-only ring size 4</td></tr> <tr><td>134</td><td>>= 1 unsaturated non-aromatic nitrogen-containing ring size 4</td></tr> <tr><td>135</td><td>>= 1 unsaturated non-aromatic heteroatom-containing ring size 4</td></tr> <tr><td>...</td><td></td></tr> </tbody> </table>	Bit Position	Bit Substructure	115	>= 1 any ring size 3	116	>= 1 saturated or aromatic carbon-only ring size 3	117	>= 1 saturated or aromatic nitrogen-containing ring size 3	118	>= 1 saturated or aromatic heteroatom-containing ring size 3	119	>= 1 unsaturated non-aromatic carbon-only ring size 3	120	>= 1 unsaturated non-aromatic nitrogen-containing ring size 3	121	>= 1 unsaturated non-aromatic heteroatom-containing ring size 3	...		129	>= 1 any ring size 4	130	>= 1 saturated or aromatic carbon-only ring size 4	131	>= 1 saturated or aromatic nitrogen-containing ring size 4	132	>= 1 saturated or aromatic heteroatom-containing ring size 4	133	>= 1 unsaturated non-aromatic carbon-only ring size 4	134	>= 1 unsaturated non-aromatic nitrogen-containing ring size 4	135	>= 1 unsaturated non-aromatic heteroatom-containing ring size 4	...																																																																																																																				
Position	Bit-Substructure																																																																																																																																																																																			
0	>= 4 H																																																																																																																																																																																			
1	>= 8 H																																																																																																																																																																																			
2	>= 16 H																																																																																																																																																																																			
3	>= 32 H																																																																																																																																																																																			
4	>= 1 Li																																																																																																																																																																																			
5	>= 2 Li																																																																																																																																																																																			
6	>= 1 B																																																																																																																																																																																			
7	>= 2 B																																																																																																																																																																																			
8	>= 4 B																																																																																																																																																																																			
9	>= 2 C																																																																																																																																																																																			
10	>= 4 C																																																																																																																																																																																			
11	>= 8 C																																																																																																																																																																																			
12	>= 16 C																																																																																																																																																																																			
...																																																																																																																																																																																				
Bit Position	Bit Substructure																																																																																																																																																																																			
115	>= 1 any ring size 3																																																																																																																																																																																			
116	>= 1 saturated or aromatic carbon-only ring size 3																																																																																																																																																																																			
117	>= 1 saturated or aromatic nitrogen-containing ring size 3																																																																																																																																																																																			
118	>= 1 saturated or aromatic heteroatom-containing ring size 3																																																																																																																																																																																			
119	>= 1 unsaturated non-aromatic carbon-only ring size 3																																																																																																																																																																																			
120	>= 1 unsaturated non-aromatic nitrogen-containing ring size 3																																																																																																																																																																																			
121	>= 1 unsaturated non-aromatic heteroatom-containing ring size 3																																																																																																																																																																																			
...																																																																																																																																																																																				
129	>= 1 any ring size 4																																																																																																																																																																																			
130	>= 1 saturated or aromatic carbon-only ring size 4																																																																																																																																																																																			
131	>= 1 saturated or aromatic nitrogen-containing ring size 4																																																																																																																																																																																			
132	>= 1 saturated or aromatic heteroatom-containing ring size 4																																																																																																																																																																																			
133	>= 1 unsaturated non-aromatic carbon-only ring size 4																																																																																																																																																																																			
134	>= 1 unsaturated non-aromatic nitrogen-containing ring size 4																																																																																																																																																																																			
135	>= 1 unsaturated non-aromatic heteroatom-containing ring size 4																																																																																																																																																																																			
...																																																																																																																																																																																				
<p>Section 3: Simple atom pairs - These bits test for the presence of patterns of bonded atom pairs, regardless of bond order or count.</p> <table border="1"> <tbody> <tr><td>263</td><td>Li-H</td></tr> <tr><td>264</td><td>Li-Li</td></tr> <tr><td>265</td><td>Li-B</td></tr> <tr><td>266</td><td>Li-C</td></tr> <tr><td>267</td><td>Li-O</td></tr> <tr><td>268</td><td>Li-F</td></tr> <tr><td>269</td><td>Li-P</td></tr> <tr><td>270</td><td>Li-S</td></tr> <tr><td>271</td><td>Li-Cl</td></tr> <tr><td>272</td><td>B-H</td></tr> <tr><td>273</td><td>B-B</td></tr> <tr><td>274</td><td>B-C</td></tr> <tr><td>275</td><td>B-N</td></tr> <tr><td>276</td><td>B-O</td></tr> <tr><td>277</td><td>B-F</td></tr> <tr><td>278</td><td>B-Si</td></tr> </tbody> </table>	263	Li-H	264	Li-Li	265	Li-B	266	Li-C	267	Li-O	268	Li-F	269	Li-P	270	Li-S	271	Li-Cl	272	B-H	273	B-B	274	B-C	275	B-N	276	B-O	277	B-F	278	B-Si	<p>Section 4: Simple atom nearest neighbors - These bits test for the presence of atom nearest neighbor patterns, regardless of bond order (denoted by "~") or count, but where bond aromaticity (denoted by ":") is significant.</p> <table border="1"> <tbody> <tr><td>327</td><td>C(~Br) (~C)</td></tr> <tr><td>328</td><td>C(~Br) (~C) (~C)</td></tr> <tr><td>329</td><td>C(~Br) (~H)</td></tr> <tr><td>330</td><td>C(~Br) (:C)</td></tr> <tr><td>331</td><td>C(~Br) (:N)</td></tr> <tr><td>332</td><td>C(~C) (~C)</td></tr> <tr><td>...</td><td></td></tr> <tr><td>629</td><td>S-C:C:C-N</td></tr> <tr><td>630</td><td>O-C:C-O-C</td></tr> <tr><td>631</td><td>O-C:C-O-[#1]</td></tr> <tr><td>632</td><td>C-C-O-C:C</td></tr> <tr><td>633</td><td>N-C-C:C-C</td></tr> <tr><td>634</td><td>C-C-C:C-C</td></tr> <tr><td>635</td><td>N-N-C-N-[#1]</td></tr> <tr><td>...</td><td></td></tr> <tr><td>709</td><td>C-C(C)-O-C-C</td></tr> <tr><td>710</td><td>C-C(C)(C)-C-C</td></tr> <tr><td>711</td><td>C-C(C)(C)-O-C-C</td></tr> <tr><td>712</td><td>C-C(C)-C(C)-C-C</td></tr> </tbody> </table>	327	C(~Br) (~C)	328	C(~Br) (~C) (~C)	329	C(~Br) (~H)	330	C(~Br) (:C)	331	C(~Br) (:N)	332	C(~C) (~C)	...		629	S-C:C:C-N	630	O-C:C-O-C	631	O-C:C-O-[#1]	632	C-C-O-C:C	633	N-C-C:C-C	634	C-C-C:C-C	635	N-N-C-N-[#1]	...		709	C-C(C)-O-C-C	710	C-C(C)(C)-C-C	711	C-C(C)(C)-O-C-C	712	C-C(C)-C(C)-C-C	<p>Section 5: Detailed atom neighborhoods - These bits test for the presence of detailed atom neighborhood patterns, regardless of count, but where bond orders are specific, bond aromaticity matches both single and double bonds, and where "-", "=", and "#" matches a single bond, double bond, and triple bond order, respectively.</p> <table border="1"> <tbody> <tr><td>416</td><td>C=C</td></tr> <tr><td>417</td><td>C#C</td></tr> <tr><td>418</td><td>C=N</td></tr> <tr><td>419</td><td>C#N</td></tr> <tr><td>420</td><td>C=O</td></tr> <tr><td>421</td><td>C=S</td></tr> <tr><td>422</td><td>N=N</td></tr> <tr><td>423</td><td>N=O</td></tr> <tr><td>424</td><td>N=P</td></tr> <tr><td>425</td><td>P=O</td></tr> <tr><td>426</td><td>P=P</td></tr> <tr><td>427</td><td>C(#C) (-C)</td></tr> <tr><td>428</td><td>C(#C) (-H)</td></tr> <tr><td>429</td><td>C(#N) (-C)</td></tr> </tbody> </table>	416	C=C	417	C#C	418	C=N	419	C#N	420	C=O	421	C=S	422	N=N	423	N=O	424	N=P	425	P=O	426	P=P	427	C(#C) (-C)	428	C(#C) (-H)	429	C(#N) (-C)	<p>Section 6: Simple SMARTS patterns - These bits test for the presence of simple SMARTS patterns, regardless of count, but where bond orders are specific and bond aromaticity matches both single and double bonds.</p> <table border="1"> <tbody> <tr><td>460</td><td>C-C-C#C</td></tr> <tr><td>461</td><td>O-C-C=N</td></tr> <tr><td>462</td><td>O-C-C=O</td></tr> <tr><td>463</td><td>N:C-S-[#1]</td></tr> <tr><td>464</td><td>N-C-C=C</td></tr> <tr><td>465</td><td>O=S-C-C</td></tr> <tr><td>466</td><td>N#C-C=C</td></tr> <tr><td>467</td><td>C=N-N-C</td></tr> <tr><td>468</td><td>O=S-C-N</td></tr> <tr><td>469</td><td>S-S-C:C</td></tr> <tr><td>470</td><td>C:C-C=C</td></tr> <tr><td>471</td><td>S:C:C:C</td></tr> <tr><td>472</td><td>C:N:C-C</td></tr> <tr><td>473</td><td>S-C:N:C</td></tr> <tr><td>474</td><td>S:C:C:N</td></tr> <tr><td>475</td><td>S-C=N-C</td></tr> </tbody> </table>	460	C-C-C#C	461	O-C-C=N	462	O-C-C=O	463	N:C-S-[#1]	464	N-C-C=C	465	O=S-C-C	466	N#C-C=C	467	C=N-N-C	468	O=S-C-N	469	S-S-C:C	470	C:C-C=C	471	S:C:C:C	472	C:N:C-C	473	S-C:N:C	474	S:C:C:N	475	S-C=N-C	<p>Section 7: Complex SMARTS patterns - These bits test for the presence of complex SMARTS patterns, regardless of count, but where bond orders and bond aromaticity are specific.</p> <table border="1"> <tbody> <tr><td>713</td><td>Cc1ccc(C)cc1</td></tr> <tr><td>714</td><td>Cc1ccc(O)cc1</td></tr> <tr><td>715</td><td>Cc1ccc(S)cc1</td></tr> <tr><td>716</td><td>Cc1ccc(N)cc1</td></tr> <tr><td>717</td><td>Cc1ccc(Cl)cc1</td></tr> <tr><td>718</td><td>Cc1ccc(Br)cc1</td></tr> <tr><td>719</td><td>Oc1ccc(O)cc1</td></tr> <tr><td>720</td><td>Oc1ccc(S)cc1</td></tr> <tr><td>721</td><td>Oc1ccc(N)cc1</td></tr> <tr><td>722</td><td>Oc1ccc(Cl)cc1</td></tr> <tr><td>723</td><td>Oc1ccc(Br)cc1</td></tr> <tr><td>724</td><td>Sc1ccc(S)cc1</td></tr> <tr><td>725</td><td>Sc1ccc(N)cc1</td></tr> <tr><td>726</td><td>Sc1ccc(Cl)cc1</td></tr> <tr><td>727</td><td>Sc1ccc(Br)cc1</td></tr> <tr><td>728</td><td>Nc1ccc(N)cc1</td></tr> <tr><td>729</td><td>Nc1ccc(Cl)cc1</td></tr> <tr><td>730</td><td>Nc1ccc(Br)cc1</td></tr> <tr><td>731</td><td>Clc1ccc(Cl)cc1</td></tr> <tr><td>732</td><td>Clc1ccc(Br)cc1</td></tr> <tr><td>733</td><td>Brclccc(Br)cc1</td></tr> <tr><td>734</td><td>Cc1cc(C)ccc1</td></tr> <tr><td>...</td><td></td></tr> </tbody> </table>	713	Cc1ccc(C)cc1	714	Cc1ccc(O)cc1	715	Cc1ccc(S)cc1	716	Cc1ccc(N)cc1	717	Cc1ccc(Cl)cc1	718	Cc1ccc(Br)cc1	719	Oc1ccc(O)cc1	720	Oc1ccc(S)cc1	721	Oc1ccc(N)cc1	722	Oc1ccc(Cl)cc1	723	Oc1ccc(Br)cc1	724	Sc1ccc(S)cc1	725	Sc1ccc(N)cc1	726	Sc1ccc(Cl)cc1	727	Sc1ccc(Br)cc1	728	Nc1ccc(N)cc1	729	Nc1ccc(Cl)cc1	730	Nc1ccc(Br)cc1	731	Clc1ccc(Cl)cc1	732	Clc1ccc(Br)cc1	733	Brclccc(Br)cc1	734	Cc1cc(C)ccc1	...	
263	Li-H																																																																																																																																																																																			
264	Li-Li																																																																																																																																																																																			
265	Li-B																																																																																																																																																																																			
266	Li-C																																																																																																																																																																																			
267	Li-O																																																																																																																																																																																			
268	Li-F																																																																																																																																																																																			
269	Li-P																																																																																																																																																																																			
270	Li-S																																																																																																																																																																																			
271	Li-Cl																																																																																																																																																																																			
272	B-H																																																																																																																																																																																			
273	B-B																																																																																																																																																																																			
274	B-C																																																																																																																																																																																			
275	B-N																																																																																																																																																																																			
276	B-O																																																																																																																																																																																			
277	B-F																																																																																																																																																																																			
278	B-Si																																																																																																																																																																																			
327	C(~Br) (~C)																																																																																																																																																																																			
328	C(~Br) (~C) (~C)																																																																																																																																																																																			
329	C(~Br) (~H)																																																																																																																																																																																			
330	C(~Br) (:C)																																																																																																																																																																																			
331	C(~Br) (:N)																																																																																																																																																																																			
332	C(~C) (~C)																																																																																																																																																																																			
...																																																																																																																																																																																				
629	S-C:C:C-N																																																																																																																																																																																			
630	O-C:C-O-C																																																																																																																																																																																			
631	O-C:C-O-[#1]																																																																																																																																																																																			
632	C-C-O-C:C																																																																																																																																																																																			
633	N-C-C:C-C																																																																																																																																																																																			
634	C-C-C:C-C																																																																																																																																																																																			
635	N-N-C-N-[#1]																																																																																																																																																																																			
...																																																																																																																																																																																				
709	C-C(C)-O-C-C																																																																																																																																																																																			
710	C-C(C)(C)-C-C																																																																																																																																																																																			
711	C-C(C)(C)-O-C-C																																																																																																																																																																																			
712	C-C(C)-C(C)-C-C																																																																																																																																																																																			
416	C=C																																																																																																																																																																																			
417	C#C																																																																																																																																																																																			
418	C=N																																																																																																																																																																																			
419	C#N																																																																																																																																																																																			
420	C=O																																																																																																																																																																																			
421	C=S																																																																																																																																																																																			
422	N=N																																																																																																																																																																																			
423	N=O																																																																																																																																																																																			
424	N=P																																																																																																																																																																																			
425	P=O																																																																																																																																																																																			
426	P=P																																																																																																																																																																																			
427	C(#C) (-C)																																																																																																																																																																																			
428	C(#C) (-H)																																																																																																																																																																																			
429	C(#N) (-C)																																																																																																																																																																																			
460	C-C-C#C																																																																																																																																																																																			
461	O-C-C=N																																																																																																																																																																																			
462	O-C-C=O																																																																																																																																																																																			
463	N:C-S-[#1]																																																																																																																																																																																			
464	N-C-C=C																																																																																																																																																																																			
465	O=S-C-C																																																																																																																																																																																			
466	N#C-C=C																																																																																																																																																																																			
467	C=N-N-C																																																																																																																																																																																			
468	O=S-C-N																																																																																																																																																																																			
469	S-S-C:C																																																																																																																																																																																			
470	C:C-C=C																																																																																																																																																																																			
471	S:C:C:C																																																																																																																																																																																			
472	C:N:C-C																																																																																																																																																																																			
473	S-C:N:C																																																																																																																																																																																			
474	S:C:C:N																																																																																																																																																																																			
475	S-C=N-C																																																																																																																																																																																			
713	Cc1ccc(C)cc1																																																																																																																																																																																			
714	Cc1ccc(O)cc1																																																																																																																																																																																			
715	Cc1ccc(S)cc1																																																																																																																																																																																			
716	Cc1ccc(N)cc1																																																																																																																																																																																			
717	Cc1ccc(Cl)cc1																																																																																																																																																																																			
718	Cc1ccc(Br)cc1																																																																																																																																																																																			
719	Oc1ccc(O)cc1																																																																																																																																																																																			
720	Oc1ccc(S)cc1																																																																																																																																																																																			
721	Oc1ccc(N)cc1																																																																																																																																																																																			
722	Oc1ccc(Cl)cc1																																																																																																																																																																																			
723	Oc1ccc(Br)cc1																																																																																																																																																																																			
724	Sc1ccc(S)cc1																																																																																																																																																																																			
725	Sc1ccc(N)cc1																																																																																																																																																																																			
726	Sc1ccc(Cl)cc1																																																																																																																																																																																			
727	Sc1ccc(Br)cc1																																																																																																																																																																																			
728	Nc1ccc(N)cc1																																																																																																																																																																																			
729	Nc1ccc(Cl)cc1																																																																																																																																																																																			
730	Nc1ccc(Br)cc1																																																																																																																																																																																			
731	Clc1ccc(Cl)cc1																																																																																																																																																																																			
732	Clc1ccc(Br)cc1																																																																																																																																																																																			
733	Brclccc(Br)cc1																																																																																																																																																																																			
734	Cc1cc(C)ccc1																																																																																																																																																																																			
...																																																																																																																																																																																				

Table showing the different sections found in the PubChem Substructure Fingerprint V1.3.Extracted from [15].

EXAMPLES

ACETIC ACID

It consists of two small functional groups, a methyl group and a carboxyl group linked. Alternatively, it could be named as an acetyl and a hydroxyl group. Carbon from methyl group is a sp^3 carbon with tetrahedral geometry. Carbon in the acetyl group is a sp^2 carbon having a double bond with one of the oxygen atoms and the other being negatively charged, and then it is planar. Furthermore, the missing electron in this charged group is delocalized. This molecule has some flexibility given by the only degree of freedom that is the single bond between the two carbons.

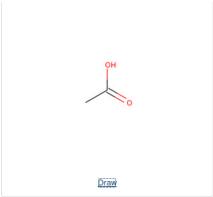
ZINC5224164 (Acetate)
In: anodyne bb for-sale in-man-only in-stock nonhuman-metabolites
Google Wikipedia PubMed

Added	Available	Since	Mwt	logP	Heavy Atoms	Tranche	Download
2006-01-26	In-Stock	2015-08-07	60.052	0.091	4	ACAA	

SMILES: CC(=O)O
InChI: InChI=1S/C2H4O2(=1-2)(3)/4(1)H3,(H,3,4)
InChI Key: QTBBSXVTEAMEQO-UHFFFAOYSA-N

Available 3D Representations Find Decoy

pH range	Net charge	H-bond donors	H-bond acceptors	TPSA	Rotatable bonds	Apolar desolvation	Polar desolvation	Download
Reference	-1	0	2	40	0	2.08	-42.29	



Top, ZINC15 acetate entry [16]. Bottom, three dimensional representation of an acetic acid molecule.

ASPIRIN

A carboxylic acid and ester groups form aspirin with only one sp^3 carbon having tetrahedral geometry (end of the carbon), all the other carbons are linked by double bonds that makes them planar. This molecule could be considered more flexible than acetic acid since it has more rotatable bonds (i.e., 4 degrees of freedom vs. 1 rotatable bond of the acetic acid).

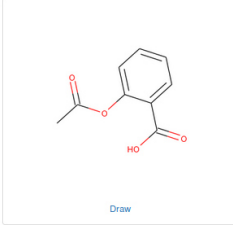
ZINC53 (Aspirin)
In: anodyne bb fda for-sale in-stock natural-products
Google Wikipedia PubMed

Added	Available	Since	Mwt	logP	Heavy Atoms	Tranche	Download
2005-09-27	In-Stock	2015-08-07	180.159	1.31	13	ADAA	

SMILES: CC(=O)Oc1ccccc1C(=O)O
InChI: InChI=1S/C9H8O4(=1-6)(10)13-9-5-3-2-4-7(8)(11)12/h2-5H,1H3,(H,11,12)
InChI Key: BSYNRYMUTXBXSQ-UHFFFAOYSA-N

Available 3D Representations Find Decoy

pH range	Net charge	H-bond donors	H-bond acceptors	TPSA	Rotatable bonds	Apolar desolvation	Polar desolvation	Download
Reference	-1	0	4	66	2	6.58	-56.82	



Top, ZINC15 aspirin entry [16]. Bottom, three-dimensional representation of an acetic acid molecule.

GUANOSINE TRIPHOSPHATE (GTP)

It plays a role in several biological processes. GTP is involved in energy transfer within the cell. For instance, one of the enzymes in the citric acid cycle generates a GTP molecule. This is equivalent to the generation of one molecule of ATP, since GTP is readily converted to ATP with nucleoside-diphosphate kinase (NDK). During the elongation stage of genetic translation, GTP is used as an energy source for the binding of a new amino-bound tRNA to the A site of the ribosome. GTP is also used as an energy source for the translocation of the ribosome towards the 3' end of the mRNA. During microtubule polymerization, each heterodimer formed by an α and a β tubulin molecule carries two GTP molecules, and the GTP is hydrolyzed to GDP when the tubulin dimers are added to the plus end of the growing microtubule. The translocation of proteins into the mitochondria matrix involves the interactions of both GTP and ATP. The importance of these proteins plays crucial role in several pathways regulated within the mitochondria organelle.

ZINC60094177 (Gtg)

In: anodyne bb endogenous for-sale in-man-only in-stock

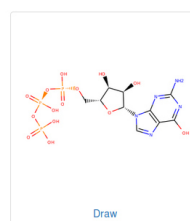
Google Wikipedia PubMed

Added	Available	Since	Mwt	logP	Heavy Atoms	Tranche	Download
2011-03-14	In-Stock	2015-08-08	523.181	-1.923	32	KAAA	Download

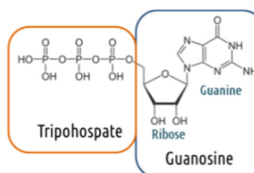
SMILES Nc1nc(O)c2ncn([C@@H]3O[C@H](CO[P@@](=O)(O)OP(=O)(O)OP(=O)(O)C@@H(O)C@H3O)c2n1
InChI InChI=1S/C10H16N5O14P3/c11-10-13-7-4(8(18)14-10)12-2-15(7)9-6(17)5(16)3(27-9)1-26-31(22,23)29-32(24,25)1
InChI Key XKMLYUALXHKNFT-UUOKFMHZSA-N

Available 3D Representations

pH range	Net charge	H-bond donors	H-bond acceptors	tPSA	Rotatable bonds	Apolar desolvation	Polar desolvation	Download
Reference	-4	4	18	310	8	-3.18	-375.18	Download
Mid pH (7.4)	-3	5	17	307	8	-4.33	-237.59	Download

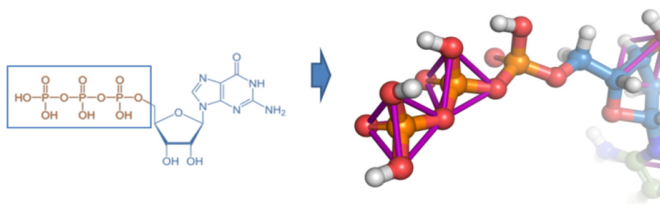


Guanosine Triphosphate

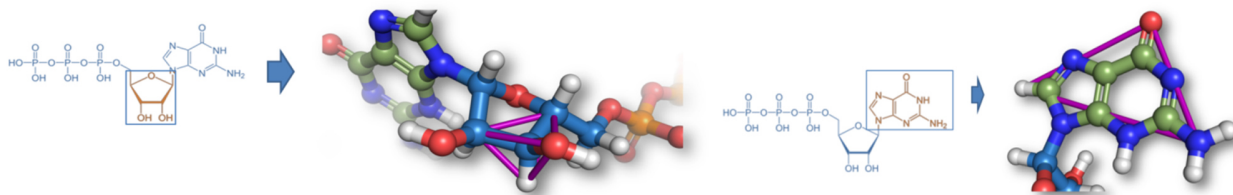


Top, ZINC15 Gtg entry [16]. Bottom, three dimensional representation of an acetic acid molecule.

In the 2D structure of the molecule, we can observe three phosphate groups linked to a nucleoside. The triphosphate group is formed by a chain of tetrahedral phosphorus atoms occupying a significant volume, and is highly flexible since it has several single bonds (i.e., rotatable bonds/degrees of freedom).



At the right of the molecule, we can observe a ribose, which has two flexible hydroxyl groups that allows the orientation of the hydrogen atoms to form hydrogen bond interactions with other molecules. A guanine nitrogenous base that is a pyrimidine-imidazole ring system with conjugated double bonds is planar.



MOLECULAR FORCE FIELDS

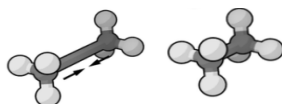
ATOMIC DETAIL COMPUTER SIMULATION

In the context of molecular modelling, a force field refers to the parameters and mathematical functions used to describe the potential energy of a system of particles (typically molecules and atoms) [17].

"All-atom" force fields provide parameters for every type of atom in a system, including hydrogen. "Coarse-grained" force fields, which are frequently used in long-time simulations of big molecular systems, provide even more crude representations for increased computational efficiency.

A force field can be used to optimize the geometry of molecules or to estimate motions of molecules. "All-atom" force fields can be used for adjusting bond lengths and angles of a molecule, looking for the minimum energy of the system. "Coarse-grained" force fields can be used for estimating the global fluctuations of a macromolecule considering the whole structure as an oscillatory system.

In this section, we will use molecular mechanics focusing on the classical definition of the internal energy of a molecule, and its application to molecular energy minimization. Then we will refer to this type of force fields as Molecular Mechanics Force Fields (MMFF).



Left, ethane molecule with carbons-carbon bond length far from its minimum energy. Right, ethane molecule with carbons-carbon bond length near its minimum energy. Image from Wikimedia Commons [7] (CC BY-NC-ND 3.0).

THE POTENTIAL ENERGY OF MOLECULES

To describe the potential energy of a molecule (or set of molecules), we will need to define an energy function that will require many parameters. The parameters can be derived from experimental work or by quantum mechanical calculations.

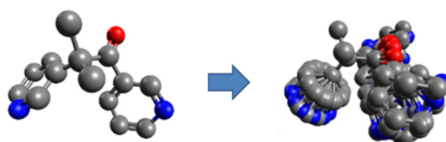
All interactions contribute to the stability of a molecule, defining the energy of a molecule as a summation of internal and non-bonded interactions,

$$E = E_{bond} + E_{Non-Bonded}$$

where bonded and non-bonded energies are defined as a summation of different terms. They account for the forces participating in intramolecular and intermolecular interactions, as bond stretching, bending or torsion, also called internal energy, or non-bonded interaction, as electrostatic and Van der Waals interactions among others

THE INTERNAL ENERGY

The internal energy of a molecule, or bonding energy, is described by the addition of terms that will keep the correct geometry of the of atoms linked together by covalent bonds. Because of high-energy constants in bond length and angle potentials, these degrees of freedom are effectively frozen at room temperature. In contrast, the dihedral angle potential of single bonds have energy constants comparable with the energy of thermal motion $k_B T$, and they determine the structural transitions in molecules, i.e. they determine the conformations of molecule. Then, the “unfrozen” potentials when looking at bonding energy is the dihedral potential.



Representation of the conformational space of a molecule by the rotation of its dihedral angles. Image adapted from Wikimedia Commons [7](CC BY-NC-ND 3.0).

Although more expressions are often included, we will focus in describing the potential function of bond stretching, angle bending and dihedral angle forces.

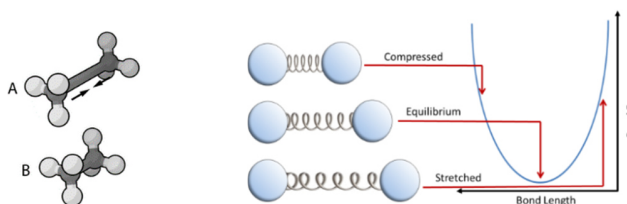
$$E_{bond} = E_{Stretching} + E_{Angle} + E_{Dihedral}$$

BOND STRETCHING POTENTIAL

To estimate the potential energy associated with the length of covalent bond, called stretching energy, we can use the Hooke's law. This equation will give an energy penalty when a certain bond length is either shorter or longer than expected, i.e. when displaced from their minimum of energy. Note that this energy term involves two atoms. Being r_{ij} the distance between the atoms i and j , r_0 the ideal bond distance between these atoms and k_{ij} the force constant, we can define de potential energy term accounting for the stretching of bond length as,

$$E_{bond} = \frac{1}{2} k_{ij} (r_{ij} - r_0)^2$$

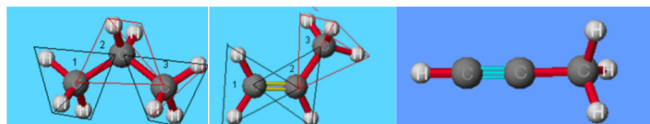
Remember that these parameters are derived either from experimental measurements or by high-level quantum calculations.



Left, representation of a bond with a distance bigger than expected (A), and a bond place at its ideal distance (B Image from Wikimedia Commons [7] (CC BY-NC-ND 3.0)). Right, representation of the potential energy function as the bond length between atoms is compressed, in equilibrium or stretched.

ANGLE BENDING POTENTIAL

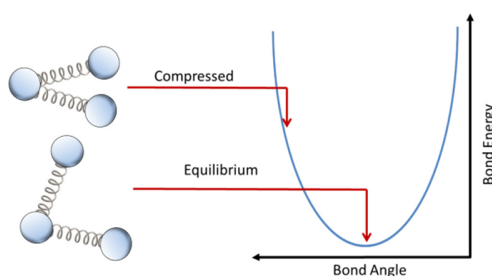
Two bonds link three consecutive atoms. These bonds form a characteristic angle for the three atoms implied in the bonds. For example, looking at the structure of the three carbon atoms forming a propane, propene or propyne molecule, we can observe that the angle between this carbon-based molecules are different attending to the geometry of the electron pair configuration of the three of atoms.



Propane, propene and propyne molecules showing different angles between the bonds of carbon atoms Image adapted from Virtual Chembook [9]

As the bond angle is bent from the norm the energy rises up, hence, we can use again the definition of the Hooke's law but instead of measuring the displacement in the bond distance, we will measure the displacement of the bond angle from its ideal position. Then, when three atoms i , j and k are linked by consecutive bonds, where the two bonds form an angle θ_{ijk} , with ideal angle between these three atoms θ_0 , and $k_{b,ijk}$ as force constant considered for each specific bond angle, we can define the potential energy of the bond angle as,

$$E_{angle} = \frac{1}{2}k_{ijk}(\theta_{ijk} - \theta_0)^2$$

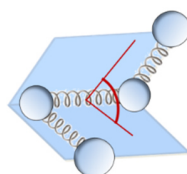


Representation of the potential energy of the bond angle between three consecutive atoms.

DIHEDRAL ANGLE POTENTIAL

The origin of dihedral angle potential, also called torsional energy, can be understood as the repulsive interactions between overlapping bond orbitals and steric clashes between atoms, and group of atoms at the sides of a certain dihedral angles of four consecutive atoms (such as C1 and C4 in butane). The potential energy of dihedral angles measures the torsion energy needed to rotate single bonds, note that double and triple bonds are rigid, and do not permit rotations.

In this energy term four consecutive atoms i , j , k and l defining two planes, and the angle between these two planes define the torsion angle.

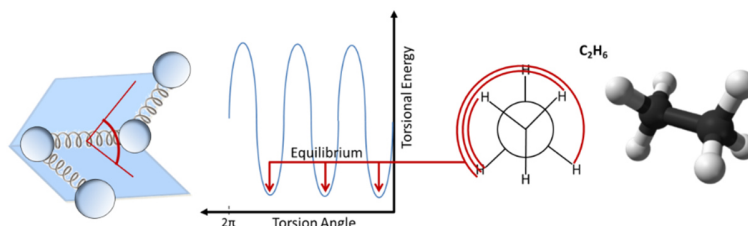


Dihedral angle defined by the two planes containing first three atoms, and second plane contain the last three atoms

A sinusoidal potential energy function can be used to model torsional energy, having one or more minimums of energy depending on the periodicity of the bond. These minimums of energy will depend on the topology of the atoms implied, as well as the chemical groups following those atoms. Taking into account these factors, the following equation defines torsional energy

$$E_{torsional} = \frac{1}{2}k_{ijkl}[1 + \cos(n\Phi_{ijkl} - \Phi_0)]$$

Where k_{ijkl} is the energy barrier, n is the periodicity of the potential, Φ_{ijkl} is the actual dihedral angle observed in the torsion bond and Φ_0 the expected dihedral angle value extracted from experimental or quantum calculations. For example, an ethane molecule will have a periodicity of three, having three minimum of energy.



Periodicity in the dihedral angle between the carbon atoms of ethane molecule. Image from Wikimedia Commons [7] (CC BY-NC-ND 3.0)

NON-BONDED ENERGY TERMS

THE VAN DER WAALS (VDW) INTERACTION

Van der Waals (vdW) interactions take place when two atoms, come into close contact. It describes two different physical phenomenon, in one hand, the repulsion due to the overlap of electron clouds performing only when the two atoms are at small distances, in the other, the attraction performing at longer distances due a phenomenon called as London's dispersion forces that explains the correlations between the atom's temporary dipole caused by the electrons position. vdW is a shape or volume concern, rather than a pure electrostatic matter.

Both forces depend inversely on the distance. Modelling them is relatively straightforward. The most widely used model is the Lenard-Jones' potential, where the repulsion term depends on r^{12} and the attractive part on r^6 , being r the separation between the atom centers.

$$E_{vdW} = \frac{A}{r^{12}} - \frac{B}{r^6}$$

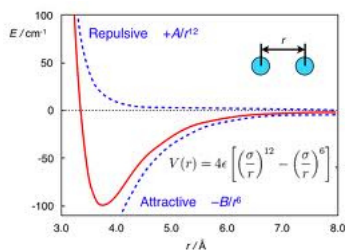
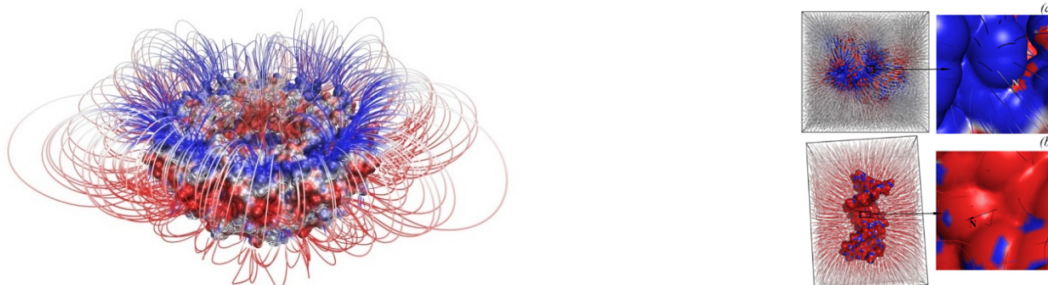


Figure showing vdW the potential between two atoms. Image from Wikimedia Commons [7] (CC BY-NC-ND 3.0).

THE ELECTROSTATIC INTERACTION

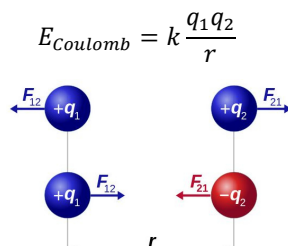
These interactions are the most ubiquitous of all, taking place in a great number of binding events, i.e. pure charge-charge interactions, hydrogen bonding, π - π stacking, hydrophobic interactions, solvation, etc. However, its accurate calculation still represents a major challenge in computational chemistry.



Left, the electrostatic field lines of TRAP---trp RNA binding attenuation protein (PDB ID: 2EXS) and right, electrostatic surface on a protein (a) and DNA (b). The blue and red colors represent the positive and negative polarities, respectively. The field lines indicate the directions of the electrostatic forces. Images from Svintradze, D. V. et al. (2017) [18]. Attribution 4.0 International (CC BY 4.0).

COULOMB POTENTIAL

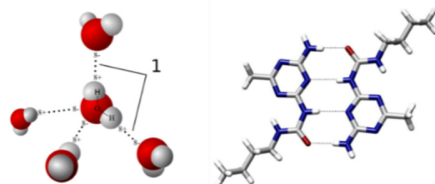
The simplest approach for evaluating pure charge-charge interactions is the Coulomb's model for the Electric potential energy, which is the product of the charges q_1 and q_2 of the atoms, over distance between them r the screened by a dielectric constant k ,



Left, Representation of the coulomb potential. Right, representation of the electrostatic potential in the surface of proteins. Image from adapted from Wikimedia Commons [7] (CC BY-NC-ND 3.0).

HYDROGEN BONDING

Hydrogen bonding is a highly selective interaction established between a hydrogen atom attached to an electronegative atom, the so-called hydrogen bond donor, and other electronegative atom, the so-called hydrogen bond acceptor. The strength of a hydrogen bond depends on the relative position of the three atoms involved, due to the preference of certain distance and angle geometries. Its role in molecular recognition is of paramount importance. As a generalization, we can range the energies associated with hydrogen bonds between 6-30 kJ/mol \approx 1.4-7kcal/mol \approx 2-12kT/e.



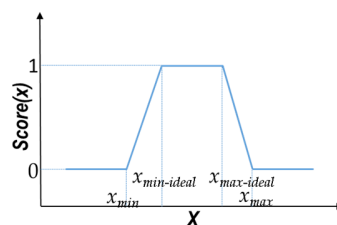
Left, the representation of a hydrogen bonds network taking place between different water molecules. Right hydrogen bond network between to molecules. Images from Wikimedia Commons [7] (CC BY-NC-ND 3.0).

Many different approaches characterize the existence of a hydrogen bond (HB). Some methods, as MM-ISMSA [19], define parameters involving the three atoms from the HB interaction (donor, D ; acceptor, A ; and the proper hydrogen atom, H), plus the atom bonded to A (X). This gives the three geometrical parameters describing the relative disposition of these atoms: first the $A-H$ distance, second the $D-H-A$ angle, and third the $H-A-X$ angle. Therefore, estimation of HB energy is performed as,

$$HB - Score(i) = \prod_x score(x), \text{ where}$$

$$score(x) = \begin{cases} 1 & \text{if } x_{min-ideal} \leq x \leq x_{max-ideal} \\ 1 - \frac{x_{min-ideal} - x}{x_{min-ideal} - x_{min}} & \text{if } x_{min} \leq x \leq x_{min-ideal} \\ 1 - \frac{x - x_{min-ideal}}{x_{max} - x_{min-ideal}} & \text{if } x_{max-ideal} \leq x \leq x_{max} \\ 0 & \text{if } x_{max} < x \\ 0 & \text{if } x < x_{min} \end{cases}$$

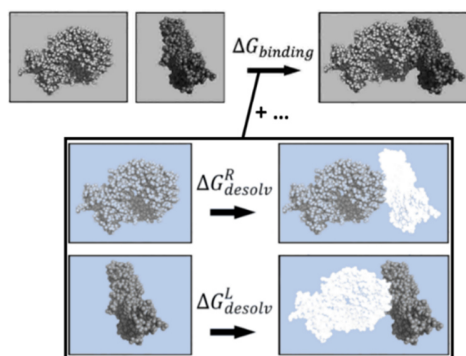
Where x represents either the distance or the angles previously described, and x_{max} , $x_{max-ideal}$, x_{min} and $x_{min-ideal}$ result from values extracted from an statistical analysis HB observed in experimental structures. This results into a function of the following form



Representation of the scoring function defined for the parameters r or the angles α and β expressed in the form of x where x_{max} , $x_{max-ideal}$, x_{min} and $x_{min-ideal}$ are obtained from the statistical values extracted from HB observed in experimental structures.

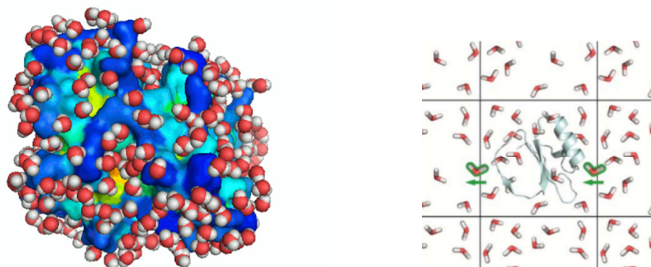
THE EFFECT OF THE SOLVENT

Solvation or desolvation energies account for the addition or removal of solvent to a molecular system. It is an important energetic effect to account for since solvent molecules mediate interactions in most biological processes. Before any molecular interaction take place, there are water molecules surrounding (solvating) the binding partners. The water molecules placed at the binding surface have to be displaced to allow the interaction. Removing this water layer for making the interaction has an energetic cost. The binding energy interaction between the molecules has to balance the energy loss from the electrostatic or Van der Waals interactions between the solvent and the molecules.



Representation of the energetic cost of desolvating two molecules before they bind.

From the modelling perspective, we can define explicit models, where the solvent molecules are represented in atomic detail and considering specifically the energy terms as electrostatic and vdW. In addition, we can define a box of waters molecules where the molecule of interest could be simulated, for example in molecular mechanics force field simulations.



Left, protein surface (represented in blue) solvated by water molecules (represented in spheres). Image from K. P. Tan et al. (2013) [20]. Right, representation of a periodic water box (water molecules represented in sticks) containing a protein (represented in light blue cartoons)

Other approaches follow implicit solvent models, where a mathematical function is built to mimic the behavior of the solvated molecular system. However, a compromise has to be reached between accuracy and speed, as these two aims are usually inversely related. Popular implicit solvent models solve the Poisson-Boltzmann equation or the Generalized Born model [21].

The Poisson-Boltzmann Model, is the classical way to deal with the electrostatic contribution to the binding energy (ΔG_{elec}), solving the PB equation using numerical methods such as finite differences, finite elements, or boundary elements. The Poisson-Boltzmann equation relates the electrostatic potential $\phi(r)$ and the charge distribution $\rho(r)$:

$$\nabla[\epsilon(r) \cdot \nabla\phi(r)] = -4\pi\rho(r)$$

where $\epsilon(r)$ is a distance-dependent dielectric function [21].

$$\Delta G_{elec}^1 = \frac{1}{2} \int \phi(r)\rho(r)dv$$

$$\Delta G_{elec}^2 = \frac{1}{2} \int \phi(r)\rho(r)dv$$

$$\Delta G_{desolv} = \Delta G_{elec}^2 - \Delta G_{elec}^1$$

Desolvation energy model, using the Poisson-Boltzmann equation. Note that the model depends on charge distribution $\phi(r)$ and the dielectric media $\rho(r)$, where r is the Cartesian coordinates of each point from the system including the molecule and the solvent.

MOLECULAR ENERGY LANDSCAPE

Having described the bonded and non-bonded energy terms we can define a Molecular Mechanics Energy Potential as the sum of the covalent and non-covalent energy terms,

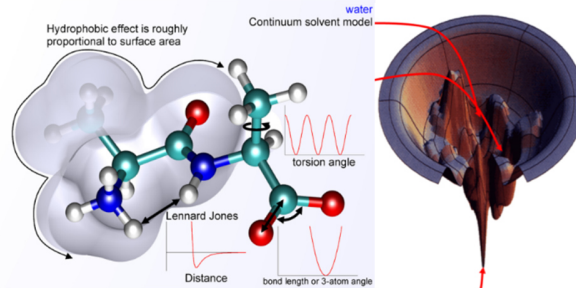
$$U(\vec{r}) = E_{bond} + E_{non-bonded}$$

where \vec{r} is the position of the atoms in the molecular system. Then, covalent energy term can be computed as the sum of bond, angle and dihedral energy terms, and non-covalent as a sum of electrostatics and van der Waals interactions among others,

$$E_{bond} = E_{stretching} + E_{angle} + E_{torsional}; E_{non-bonded} = E_{electrostatics} + E_{vanderWaals}$$

This molecular mechanics potential energy function can contain different electrostatic energy terms depending on the accuracy required in the calculation,

$$E_{electrostatics} = E_{Coulomb} + E_{hydrogenbonds} + E_{solvation} + \dots$$



Representation of the energy landscape generated by a Molecular Mechanics Potential Energy function. Images adapted from Wikimedia Commons [7] (CC BY-NC-ND 3.0)

ENTROPIC CONTRIBUTION

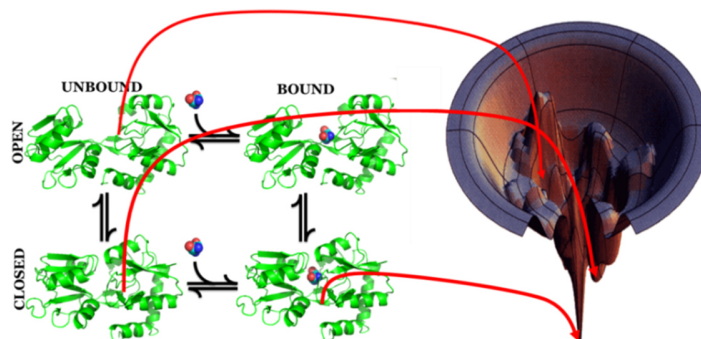
It is important to take into account that, in this definition of potential energy of a molecular system we are only representing the enthalpic contribution to the global energy of the system. Overcoming the entropic contribution involve complex calculations to sample the possible states of the system, i.e. the possible conformations of solvent and molecule atoms, which makes it expensive in computational terms.

For example, thinking in the possible conformations taking place in the binding process between two molecules, the system's entropy related to the position of its constituent particles, defines the configuration Entropy rather than to their velocity or momentum, i.e. related to the possible states of the systems.

Then, if a certain molecule has several rotatable, the number of possible states of the system, i.e. the number of possible conformations, is high and increasing the configuration entropy of the system. But in reality the theory underlying these concepts involve complex thermodynamic calculations that in many cases are difficult to incorporate them in classical computational approaches when modelling a molecular systems with thousands of atoms[22], [23].

MOLECULAR ENERGY MINIMIZATION

The molecular energy minimization process consists in the exploration of the geometries of a certain molecule towards a local minimum of energy, from the energy surface defined by a potential function. These geometries are stable along the time as long as no other physical phenomena provides the energy to push it towards another local minimum of energy. The energy required to jump from one minimum to another, should be enough to overcome the energy barrier between those two stable conformations of the molecule.



Representation of the energy landscape of a protein (green cartoons) when binding a ligand (spheres). The molecular system is associated with an energy landscape (right) with locals and global minimum of energy. A protein-ligand system can be found in open and closed conformation when it is bound or unbound to the ligand, corresponding with different points in the energy landscape.

There are several computational approaches to find these stable conformations, but most of them use iterative optimization algorithm for finding a local minimum of a differentiable function, as steepest descent or conjugate gradient minimization that will be reviewed below.

The differentiable function will be the molecular potential energy previously defined U depending on two types of variables, coordinates and parameters:

$$U = U(\text{Coordinates}; \text{Parameters})$$

The optimization process will find the coordinates minimizing the potential energy for a given set of parameters, i.e. atom types. Then, we will use the gradient of the potential energy is function of a system of N atoms, which is a vector with $3N$ components, to find the direction where the energy decrease:

$$\nabla U = \left(\frac{\partial U}{\partial x_1}, \frac{\partial U}{\partial y_1}, \frac{\partial U}{\partial z_1}, \dots, \frac{\partial U}{\partial x_N}, \frac{\partial U}{\partial y_N}, \frac{\partial U}{\partial z_N} \right)$$

Using, necessary condition for a minimum is that the function gradient is zero,

$$\nabla U = 0 \Leftrightarrow \frac{\partial U}{\partial x_i} = \frac{\partial U}{\partial y_i} = \frac{\partial U}{\partial z_i} = 0 \forall i = 1, \dots, N$$

where x_i, y_i, z_i denote atomic Cartesian coordinates and N is the number of atoms and the sufficient condition for a minimum is that the second derivative matrix \mathbf{H} is positive definite:

$$H_{ij} = \begin{pmatrix} \frac{\partial^2 U}{\partial x_1 \partial x_1} & \dots & \frac{\partial^2 U}{\partial x_1 \partial z_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 U}{\partial z_n \partial x_1} & \dots & \frac{\partial^2 U}{\partial z_n \partial z_n} \end{pmatrix}$$

We will define a convergence criteria for the iterative process to decide whether we are close enough to a certain minimum of energy

MINIMIZATION METHODS

As previously commented, minima occurs when the first derivative is zero and when the second derivative is positive. $U(\vec{r})$ is the molecular potential energy function described by Molecular Mechanics Force Fields, varying quickly with atomic coordinates \vec{r} . Then the molecular energy minimization consists in a series of steps, where the coordinates at step $n+1$ are determined from coordinates at previous step n .

$$\vec{r}_{n+1} = f(\vec{r}_n)$$

Normally, one of these two factors determines when a minimization calculation is completed:

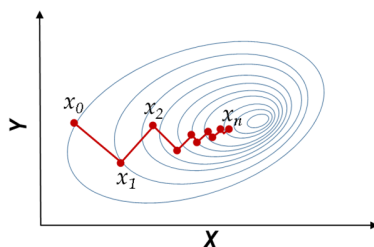
- The number of defined steps to be calculated.
- Convergence criteria, defined by a parameter that indicates when a predefined value of the gradient has been reached, that rarely reaches exactly zero value.

In steepest descent minimization method [24], the coordinates at each step are computed in the direction of fastest decrease of potential U , opposite to the gradient vector and with a certain scaling length:

$$\vec{r}_{n+1} = \vec{r}_n - \nabla U(\vec{r}_n) a$$

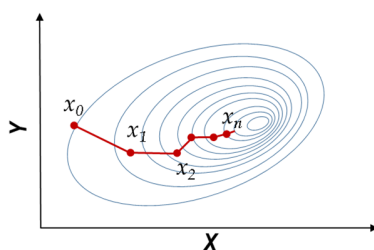
where a is a scalar parameter scaling the length of a step in the opposite direction of the gradient where the initial step is a guess.

It is not a very efficient method, but good enough for optimizing initial distorted structures. Take into account that it may be slow near a solution (i.e. minimum of energy).



Schematic representation of, the phase diagram of a potential energy function near a local minimum of energy (ellipsoids), and the “path” described by the different positions of the atoms towards the local minimum of energy (lines).

We can understand Conjugate Gradient as modification of the steepest descent minimization algorithm to increase efficiency, where current step in the direction vector is similar to previous step vectors accumulating information about the energy function from one iteration to the next.



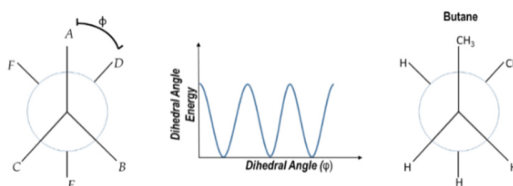
Representation of two steps of conjugate gradient energy minimization.

Some popular molecular modelling suites as Chimera [25] include minimization protocols that may combine both approaches by applying steepest descent minimization performed first to relieve highly unfavorable clashes, followed by conjugate gradient minimization, that is much slower but more effective at reaching an energy minimum after severe clashes have been relieved.

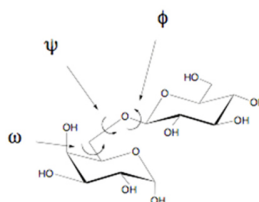
EXAMPLES

Taking into account that the first “unfrozen” degrees of freedom respect to the potential are the dihedral angles, we can represent the minimum energy by the energy of that degree of freedom, i.e. torsional energy.

- For a butane molecule, i.e. for alkyl chains, the preferred dihedral angles are 60° , 180° and 300°



- Three angles are described by ϕ , ψ and ω (in the case of glycosidic linkages via O-6). Steric considerations and anomeric effects need to be taken into consideration when looking at preferred angles. For the glycosidic bond of lipid-IVa the preferred angles are $\phi = -90^\circ$, $\psi = -100^\circ$ and $\omega = -70^\circ$.

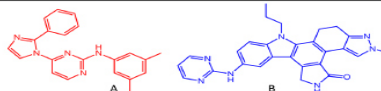
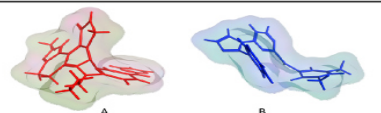
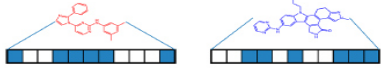



MOLECULAR SIMILARITY

When studying the structural features of molecules, it is helpful to have a measure to quantify the differences in shape between two different conformations of a molecule, or even, differences in shape or other structural properties between two different molecules.

Molecular similarity has been explored and developed concepts in chemo-informatics and medicinal chemistry. Comparing molecules and their properties, is needed medicinal chemistry research for different reasons. Searching for compounds similar to an active one, to look for compounds with free intellectual property position, or to look for compounds to complete our chemical library

Calculating similarity is a complicated problem for many reasons and can be performed at different levels, it can be evaluated based on 2D and 3D representations, and also, in terms of their physicochemical properties or their biological activity [26].

Chemical similarity	Mol. weight	LogP	Rotatable bonds	Aromatic rings	Heavy atoms
	A: 341.4	5.23	4	4	26
	B: 463.5	4.43	4	5	35
Molecular similarity					
2D similarity					
3D similarity					
Biological similarity	Vascular endothelial growth factor receptor 2		Tyrosine-protein kinase TIE-2		
	A	active	inactive		
	B	active	active		
Global similarity					
Local similarity					

Similarity perception and concepts. Two exemplary vascular endothelial growth factor receptor 2 ligands are shown, and different ways to assess their similarity are illustrated. Figure adapted with permission from F Maggiora et al. (2014) [26].

DISTANCES AND SIMILARITY MEASURES FOR SMALL MOLECULES

There are different definitions of distances, but there are some necessary conditions for them:

- $d(x, y) \in [0, \infty] \wedge d(x, y) = 0 \text{ if } \wedge \text{ only if } x = y$
- *It is symmetric:* $d(x, y) = d(y, x)$
- *It satisfies the triangle inequality:* $d(x, z) \leq d(x, y) + d(y, z)$.

In addition, there are similarity functions complementary to the definition of distance accomplishing these conditions:

- $s(x, y) \in \Lambda s(x, y) = 1 \text{ if } \wedge \text{ only if } x = y$
- *It is symmetric:* $s(x, y) = s(y, x)$
- *It satisfies the triangle inequality:* $s(x, z) \geq s(x, y) + s(y, z)$.

Most popular similarity measures or distances are based in the concept of fingerprints or in the 3D structural information of molecules.

Then, comparing fingerprints or other descriptors, we may have a discrete definition of the physicochemical properties or even fragment of the 2D structure of the molecule. Inaccurate comparisons may result in exceptional cases breaking the triangle inequality.

Methods based on the 3D coordinates would rely on geometrical properties, applying methods from algebra, geometry and physics, or even mixing them with statistical approaches for covering the diversity of the chemical structures.

Sometimes it will be more suitable to define a similarity function rather than a distance. For example, a similarity matrix composed by numbers between one and zero is often used to represent the information of the distribution of similarities over a set of elements. So, given a set of elements

$$X = \{x_1, x_2, \dots, x_n\}$$

we can define the similarity matrix S as

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{pmatrix}$$

where $s_{ij} = s(x_i, x_j)$ is the similarity found between elements x_i and x_j .

Storing this information can help to analyze differences between set of molecules as the average distance between molecules within a set, comparing distributions of distances between sets of molecules, or even applying artificial intelligence approaches as hierarchical clustering.

2D SIMILARITY, TANIMOTO COEFFICIENT

There are different ways to compare hash fingerprints, but the most used finger print similarity function in chemo informatics is called the Tanimoto coefficient [27]. It is identical to the Jaccard index as applied to binary bitstrings, and is defined as

$$f(A, B) = \frac{A \cdot B}{|A|^2 + |B|^2 - A \cdot B}$$

where,

$$A \cdot B = \sum_i (A_i \wedge B_i); |A|^2 = \sum_i (A_i^2)$$

or simply $T(A, B) = c/a + b + c$ where a and b are the number of features present in compounds A and B , respectively, and c is the number of features shared by A and B . Note that identical fingerprints have a score of 1, and when there are no bits in common the score is 0. Tanimoto score decreases if there are more bits in only one or the other fingerprint.

ROOT-MEAN-SQUARE DEVIATION (RMSD) OF ATOMIC POSITIONS

The most common definition of distance applied in molecular modelling is the Root Mean Square Deviation (RMSD). Without modifications, it will be valid for molecular systems with the same number of atoms, as for example two different conformations of the same molecule.

Then, the root-mean-square deviation (RMSD) measures the average distance between the positions of the atoms of the two molecular systems V and W ,

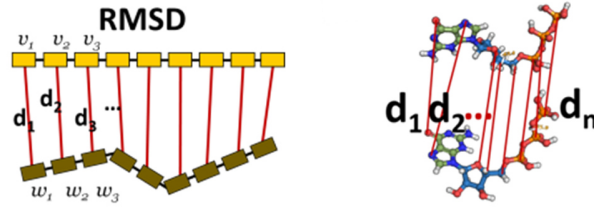
$$V = (v_1, v_2, \dots, v_n) \text{ where } v_i = (v_{ix}, v_{iy}, v_{iz})$$

$$W = (w_1, w_2, \dots, w_n) \text{ where } w_i = (w_{ix}, w_{iy}, w_{iz})$$

and it is defined as,

$$RMSD(V, W) = \sqrt{\frac{1}{n} \sum_{i=1}^n d_{eucli}^2(v_i, w_i)} = \sqrt{\frac{1}{n} \sum_{i=1}^n (v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2}$$

where (v_{ix}, v_{iy}, v_{iz}) and (w_{ix}, w_{iy}, w_{iz}) are the coordinates of the i^{th} atom and $d_{eucli}^2 = \|V - W\|^2$ the squared Euclidean distance between atom i in the different conformations v and w .



Left, schematic representation of RMSD. Right, example of RMSD between two conformations of the GTP molecules located in different places in the space.

Note that the previous definition of RMSD will compare the differences between the two molecular systems, taking into account the spatial position of the molecule under study, as a space distance. For example, it will be helpful when comparing binding models of molecular complex, as we will be indicating if the two molecules are bound to the same region of the protein or another.

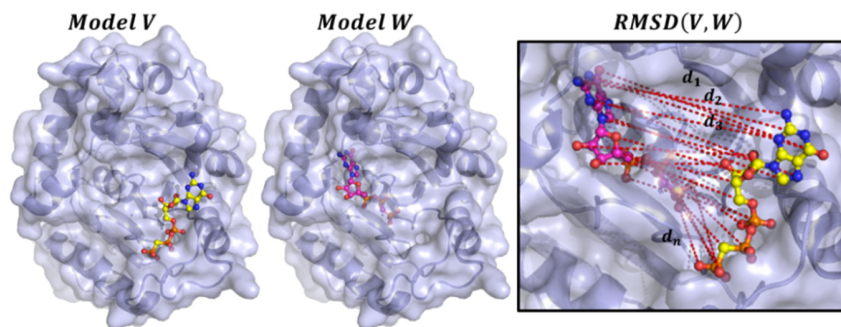


Figure displaying two models of the binding of GTP to an example protein. Note that RMSD between the atoms from the two GTP models will account for the average distance in terms of the spatial location of the models.

CRMSD: STRUCTURAL ALIGNMENT

Sometimes, it will be helpful to compare the difference between the structures of two molecules but looking at the inner differences, rather than its spatial location. For that, the common approach that applies is using a minimization algorithm to a function based on RMSD trying to lower its value by superposing (aligning) one of the structures on top of the other.

For superposing the structures, we have to use a generic spatial rigid body transformation T defined by all possible space rotations and translation:

$$T(V) = R \cdot V + M$$

Where V is the vector of the positions of the atoms of one conformation, R is the rotation matrix of a certain angle θ respect to a certain axis $u=(u_x, u_y, u_z)$ as defined as,

$$R = \begin{bmatrix} \cos\theta + u_x^2(1 - \cos\theta) & u_x u_y(1 - \cos\theta) + u_y \sin\theta & u_x u_z(1 - \cos\theta) + u_y \sin\theta \\ u_y u_x(1 - \cos\theta) + u_z \sin\theta & \cos\theta + u_y^2(1 - \cos\theta) & u_y u_z(1 - \cos\theta) + u_x \sin\theta \\ u_z u_x(1 - \cos\theta) + u_y \sin\theta & u_z u_y(1 - \cos\theta) + u_x \sin\theta & \cos\theta + u_z^2(1 - \cos\theta) \end{bmatrix}$$

and $M=(m_x, m_y, m_z)$ the translation vector.

From the definition of the RMSD, no parameters are defined as the only variables are input coordinates of the atoms from the two input conformations. However, when looking at the definition of T , it depends on the angle of rotation θ , the axis of rotation (u_x, u_y, u_z) and the translation vector (m_x, m_y, m_z) . Then, these are the seven parameters that to optimize during the minimization process in order to obtain the $cRMSD$ and obtain the structural alignment that provides the minimum RMSD.

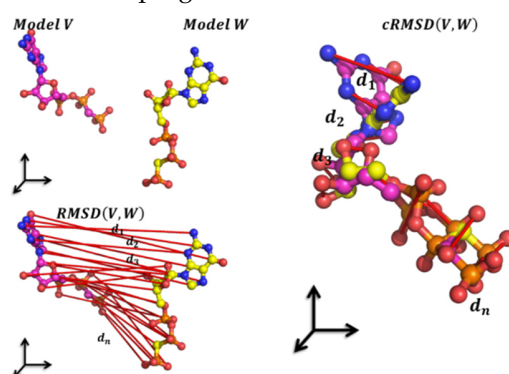
So given two conformations of the same system of atoms $V = (v_1, v_2, \dots, v_n)$ where $v_i = (v_{ix}, v_{iy}, v_{iz})$ and $W = (w_1, w_2, \dots, w_n)$ where $w_i = (w_{ix}, w_{iy}, w_{iz})$ we can define the cRMSD between proteins as,

$$cRMSD(P_A, P_B) = \min_T \left(\frac{1}{n} \sum_{i=1}^n d_{eucli}^2(v_i, T(w_i)) \right)$$

Where $T(w_i)$ is the rigid body transformation applied to W that minimizes the RMSD.

Then, when computing the cRMSD we are performing a structural alignment with the minimum RMSD between the two conformations, i.e. we are computing the optimal rigid body motion (translation and rotation) of one of the structures in order to minimize the RMSD between them.

Note that, if we save the coordinates of the conformation that has been moved to minimize that RMSD values, have aligned both molecules. That is why cRMSD can be used to obtain structural alignment between molecules in many popular molecular modelling and visualization programs.



Representation of the structural superposition/alignment of two conformations of a GTP conformation.

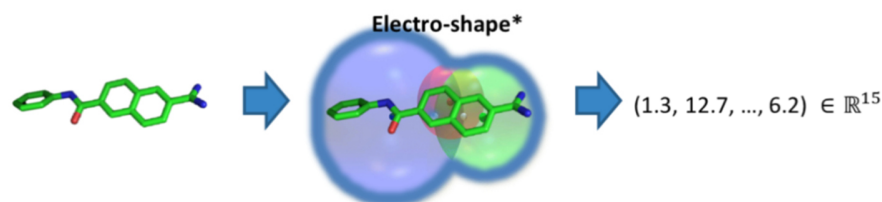
MEASURING SIMILARITY WITH ELECTROSHAPE

In order to illustrate a similarity measures used in molecular modelling we will show a fast and reliable method to compare molecules called Electro-shape [28]. Note that Electro-shape compares the space distribution of the atoms but also the charge distribution of the molecule.

The charge distribution of the molecule can be introduced by assigning partial charges to the atoms and using it as a fourth dimension. This allows us to define the structure of a molecule as a set of atoms V^{cart} defined in Cartesian coordinates and the partial charge as a fourth dimension:

$$V^{cart} = \{v_1, v_2, \dots, v_n\} = \{(x_1^{cart}, y_1^{cart}, z_1^{cart}, q_1), (x_2^{cart}, y_2^{cart}, z_2^{cart}, q_2), \dots, (x_n^{cart}, y_n^{cart}, z_n^{cart}, q_n)\} \in R^{4n}$$

For Electro-Shape similarity, we will we translate the four dimensional structure of n atoms ($\in \mathbb{R}^{4n}$) into a set of 15 Electro-Shape descriptors ($\in \mathbb{R}^{15}$).



Representation of the 15 dimension fingerprints of a sample molecule.

Then ES-descriptors are be calculated following the next steps.

First, we will compute five centroids as follows:

- the geometric center of the molecule, $C_1 = \frac{1}{n} \sum_{i=1}^n v_i = \frac{1}{n} \sum_{i=1}^n (v_{ix}, v_{iy}, v_{iz}, v_{iq})$
- the atom furthest C_1 , $C_2 = \{v_i: \max(\|v_i - C_1\|)\}$
- the atom furthest C_2 , $C_3 = \{v_i: \max(\|v_i - C_2\|)\}$
- $C_4 = C_1 + \vec{c} + (0,0,0,10^k q_{max})$
- $C_5 = C_1 + \vec{c} + (0,0,0,10^k q_{min})$

where $\vec{c} = \left(\frac{\|\vec{a}\|}{2\|\vec{a} \times \vec{b}\|} \right) (\vec{a} \times \vec{b})$, $\vec{a} = C_3 - C_1$, $\vec{b} = C_2 - C_1$ k is a scaling factor ($k=25$ suggested by the authors), q_{max} and q_{min} are the maximum and minimum partial charges q of the atoms.

Second, we will compute the distances from each atom to each of the five centroids:

$$D = \begin{pmatrix} d_{11} & \dots & d_{1n} \\ d_{21} & \dots & d_{2n} \\ d_{31} & \dots & d_{3n} \\ d_{41} & \dots & d_{4n} \\ d_{51} & \dots & d_{5n} \end{pmatrix}, \text{ where } d_{ij} = \|C_i - v_j\| \quad i \in (1, \dots, 5) \wedge j \in (1, \dots, n)$$

Finally, from the distributions of distances from the centroids to each atom D , we will compute:

- the average of the distances distribution C_1 , $es_1 = \frac{1}{n} \sum_{k=1}^n d_{1k}$
- the standard deviation of the distances distribution C_1 , $es_2 = \sqrt{\frac{1}{n} \sum_{k=1}^n (d_{1k} - es_1)^2}$
- the skewness of the distances distribution C_1 , $es_3 = \sqrt{\frac{1}{n} \sum_{k=1}^n (d_{1k} - es_1)^3}$
- the average of the distances distribution C_2 , $es_4 = \frac{1}{n} \sum_{k=1}^n d_{2k}$
- the standard deviation of the distances distribution C_2 , $es_5 = \sqrt{\frac{1}{n} \sum_{k=1}^n (d_{2k} - es_4)^2}$
- the skewness of the distances distribution C_2 , $es_6 = \sqrt{\frac{1}{n} \sum_{k=1}^n (d_{2k} - es_4)^3}$
- ...
- the average of the distances distribution C_5 , $es_{13} = \frac{1}{n} \sum_{k=1}^n d_{5k}$
- the standard deviation of the distances distribution C_5 , $es_{14} = \sqrt{\frac{1}{n} \sum_{k=1}^n (d_{5k} - es_{13})^2}$
- the skewness of the distances distribution C_5 , $es_{15} = \sqrt{\frac{1}{n} \sum_{k=1}^n (d_{5k} - es_{13})^3}$

Therefore, in this way we have transformed the $4N$ dimensions vector containing the position-charge of the atoms of the molecular system. Into a fixed 15-dimension vector that will not depend on the number the atoms on the system, nor in the spatial location of the molecule:

$$V_{ES} = \{es_1, es_2, es_3, \dots, es_{15}\} \in R^{15}$$

Having this type of descriptors, will let us compare either different conformations of the same molecule, or different molecules.

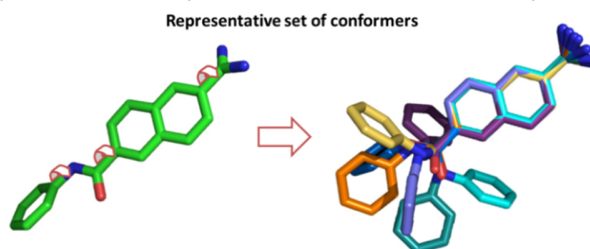
So given the Electro-shape descriptors two different molecular systems U_{ES} and V_{ES} , we can define a similarity measure that does not require any complex optimization methods as cRMSD did:

$$s_{ES}(U_{ES}, V_{ES}) = \frac{1}{1 + \sum_{i=1}^{15} |es_i^U - es_i^V|} \in (0,1]$$

Note that this approach can perform quite fast and well when comparing several molecules, but it is designed to deal specifically with small molecules, using it with macromolecules is not explored in the original publication.

CONFORMATIONAL ANALYSIS

Conformational analysis can be defined as finding a set of conformers that represents the possible states of a molecule due to its flexibility. By exploring the conformations of a molecule, we expect to obtain a set of realistic conformers, (i.e. close to a local minimum of energy) that represent the diversity of possible conformations. Conformational analysis tries to simulate this flexibility. Remember that from the definition of potential energy previously described the majority of the flexibility of a molecule is accounted by the rotations torsional bonds.

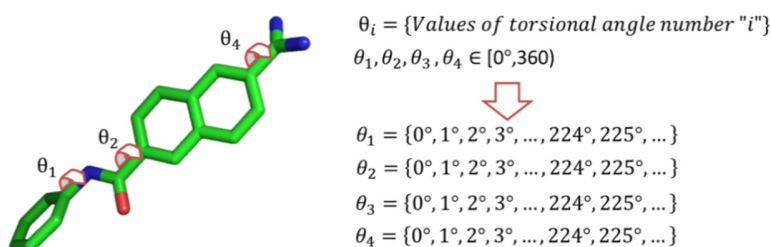


Representation of the possible conformations of a sample molecule based on its rotatable bonds.

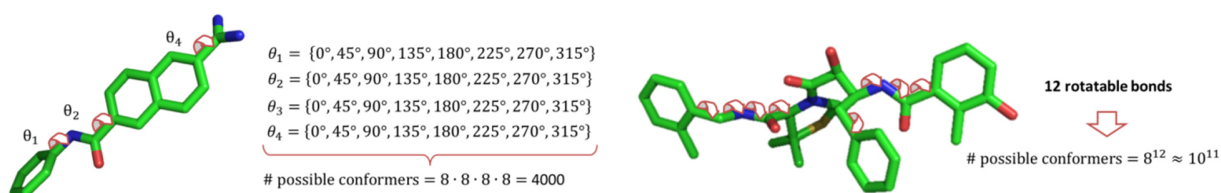
Many molecular modelling methodologies require simulating ligand flexibility as some molecular modelling approaches that will be reviewed later, protein-ligand docking, molecular dynamics, virtual screening, etc.

EXAMPLE: MODELLING MOLECULE FLEXIBILITY

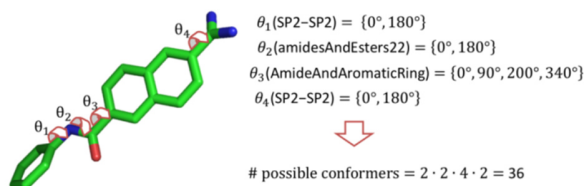
There are many different methods for exploring molecule flexibility. Next we will present an example of one of those methods ALFA (Automatic Ligand Flexibility Assignment) [29]. ALFA performs a conformational search based on angle discretization.



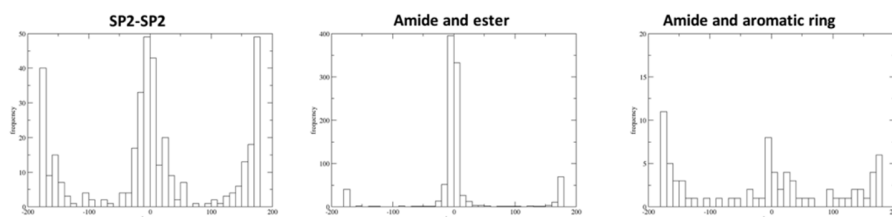
When using angle discretization we might see that, as we increase the number of rotatable bonds this simple approximation suffers and important combinatorial explosion:



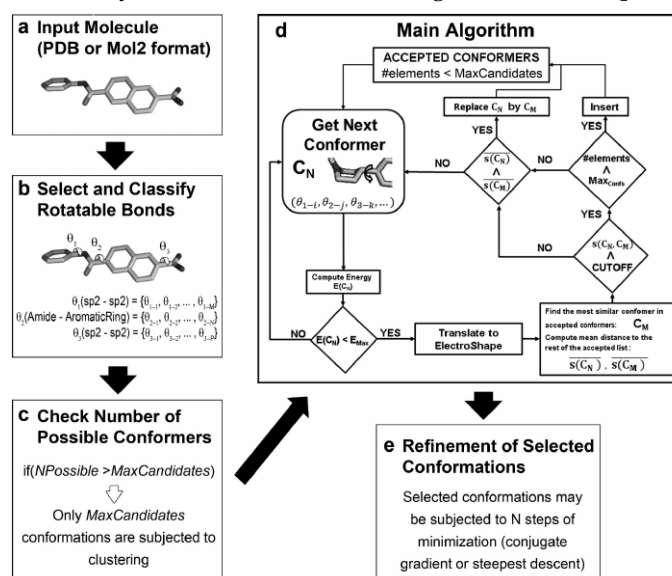
To overcome this problem, rule-based construction of conformers will reduce the amount of possible conformers selecting the angle values described by chemical rules, defining a reduced set of angles depending on bond types.



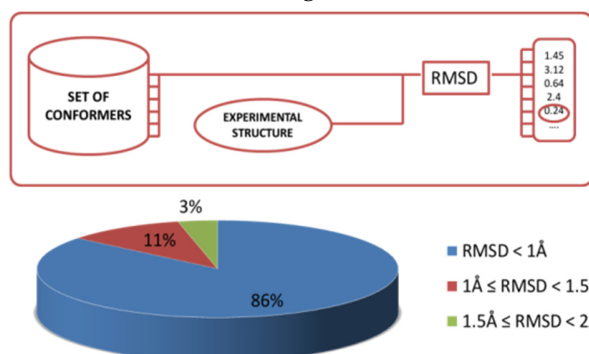
Note that if we generate conformation based on strict chemical rules, we may need to use further refinement, i.e. minimizing the selected conformations. This is because, when checking the angle values of a selected bond type over large sets of experimental structures of small molecules, most molecules follow these rules, but in some cases they fall apart from those chemical rules due to the external forces performed by interaction with other molecules. This is shown in the following histograms representing angel values for some bond types:



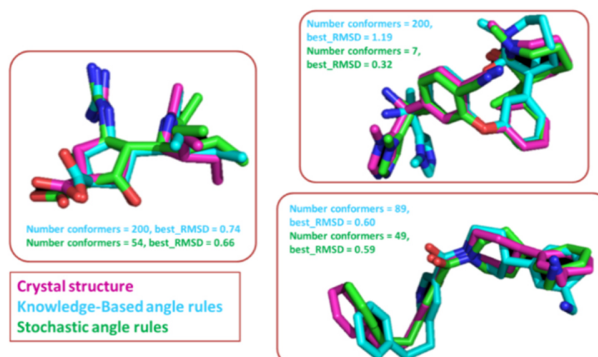
Then ALFA is computational approach that generates conformations of molecules based in knowledge-based chemical rules, Electro-shape similarity criteria, and a selection algorithm, able to perform fast in several cases.



In order to validate the method, authors checked if it is able to reproduce experimental structures of molecules. So a set of experimental structures were selected from the ASTEX^{*} data set [30], then they computed a set of conformations for each one, and measured the cRMSD between each conformer and its corresponding experimental structure. They considered that the program was performing well if it was able to produce at least one conformer bellow 1Å of cRMSD to the experimental structure, finding that this was the case for the majority of the cases.



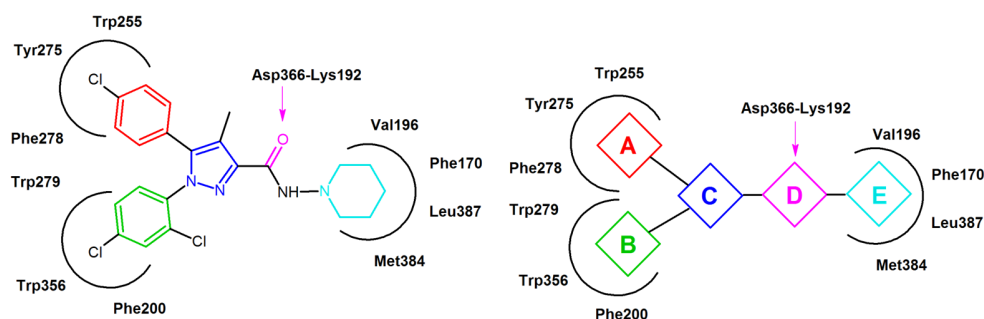
Some examples of success are shown from, where ALFA generated conformers using chemical rules and stochastic angle rules. Take these results with skepticism as they are meant just to illustrate some examples of RMSD values between small molecules.



PHARMACOPHORE MODELLING

Ligand-based pharmacophore modelling has become a key computational strategy for facilitating drug discovery in the absence of a macromolecular target structure. Then, it is used for performing similarity searches. It is carried out by extracting common chemical features from the 3D structure of ligands under study, i.e. capability to make interactions, and then search for those features in a database of ligands. Determining the essential common chemical features to construct reliable pharmacophore models should handle conformational flexibility of ligands and conducting molecular alignment. This represent the key of the technique, but also the main difficulty.

Currently, various automated pharmacophore generators have been developed, including commercially available software such as HipHop, HypoGen, DISCO, GASP, GALAHAD, PHASE and MOE; and other academic programs [31], [32].



A general C1 receptor inverse agonist pharmacophore model. Putative C1 receptor amino acid side chain residues in receptor-ligand interaction are shown. A and B both constitute an aromatic ring connected to a central core unit C. A hydrogen bond acceptor unit D interconnects unit C with a lipophilic moiety E. Rimonabant is taken as a representative example below. The applied colors indicate the mutual properties with the general C1 pharmacophore. Image from Wikimedia Commons [7] (CC BY-NC-ND 3.0)

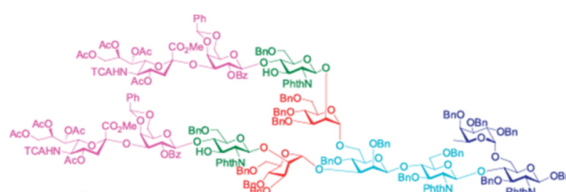
MOLECULAR SPACE

Within the set of all organic small molecules generated over the history of chemistry, one remarkable aspect is their impressive amount of them that have been already synthesized, that has remarkably increased in recent years since the implementation of combinatorial chemistry and parallel synthesis. The whole set of all synthesized molecules combining academic and commercial environments is estimated over 100 million molecules [33]. Open access database as DrugBank [34] with more than 6000 experimental or approved drugs, ChEMBL [35] with more than 1.1 million compounds with documented bioactivity, or ZINC15 [16] with more than 32 million reported molecules gathering molecular structures, physicochemical properties, and modulation activities against protein targets. Furthermore, many more molecules are theoretically possible following basic rules for covalent bonds and functional groups, so even many molecules have been already synthesized there is still a lot of room for further work [36].

THE ROLE OF SUGARS, LIPIDS AND OTHER ORGANIC MOLECULES

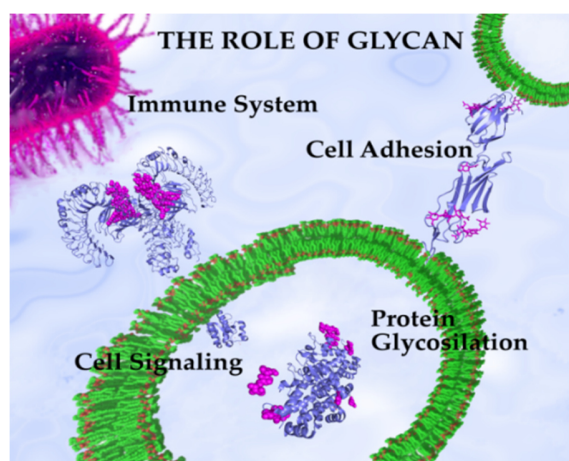
CARBOHYDRATES

Glycobiology [37] is concerned about how carbohydrates are involved in biological processes that guide life. The study of glycan (complex oligo and polysaccharides) and their interactions with other biomolecules is complicated by the inherent diversity and flexibility of these polymers, as well as by their linear and non-linear sequences.



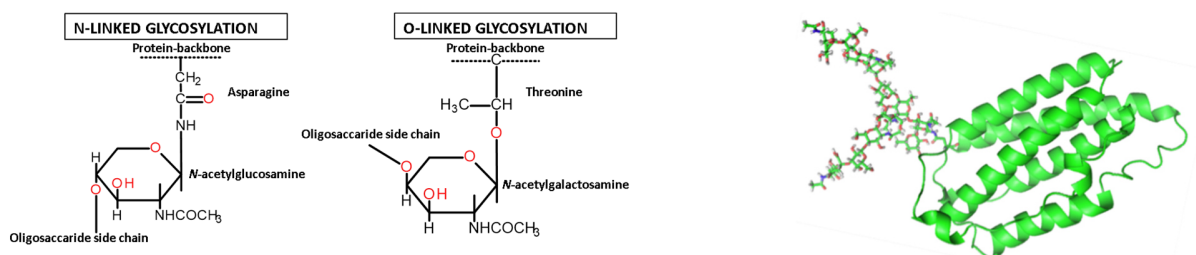
Example of a polysaccharide i.e. a carbohydrate polymer. Image from Wikimedia Commons [7] (CC BY-NC-ND 3.0)

Carbohydrates are essential in medicine, energy generation and material science. In biology, besides being the source of energy for metabolism, they are the key component of many molecular recognition processes crucial for life, as the cell signaling and adhesion, activation of the immune system, protein glycosylation, etc.



General schema of biomolecules involved in glycobiology.

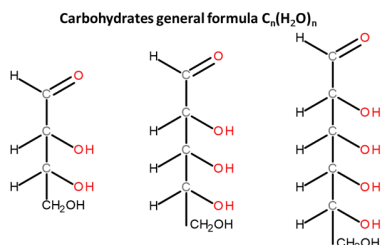
The function of glycan depends on their interactions with other molecules such as lectins, antibodies, or other entities. Many proteins in eukaryotes have glycans covalently attached, they are referred as glycoproteins [38]. There are two types of glycan bonding, O-linked glycans bonded to serine or threonine residues, or N-linked bonded through asparagine. There are also glycolipids which are glycans attached to lipids, forming the extracellular peptidoglycan or polysaccharide cell wall of bacteria. To understand how these interactions take place we have to understand the geometry of these molecules.



Right, Chemical representation of N-linked and O-linked glycosylation. Right, representation of structure of a glycosylated protein

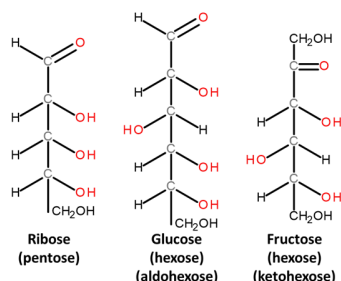
THE STRUCTURE OF CARBOHYDRATES

Carbohydrates are biomolecules accomplishing specific structural features [39]. Simple carbohydrates or monosaccharides are short hydrocarbon chains with several hydroxyls and one carbonyl functional group. Then, depending on the number of carbon atoms, they are referred as triose, tetrose, pentose, hexose, heptose, etc.



Fisher projection of 3 hydrocarbon chains. From left to right, triose, hexose, pentose and hexose

Modifications on the chirality of these hydrocarbon chains, will describe the geometry of a monosaccharide that are named following a convention, as for example, ribose, glucose, fructose or galactose.

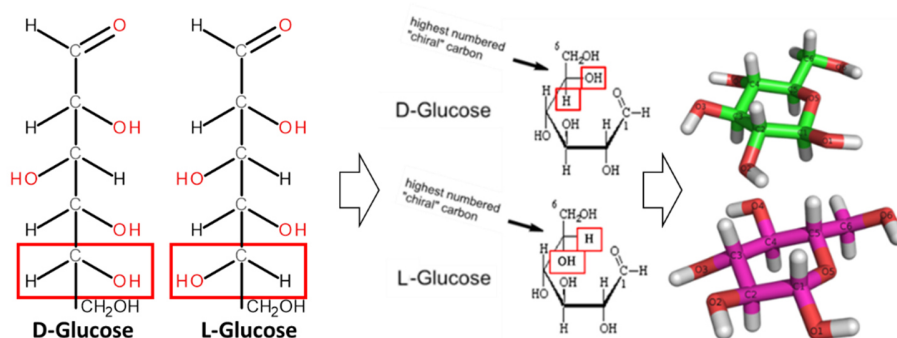


Fisher projection of three different hydrocarbon chains designated as ribose, glucose and fructose.

The cyclization of these monosaccharides occurs by a reaction between the carbonyl group or aldehyde, and the last hydroxyl group of the carbon chain. C1 or the anomeric carbon, refers to the first carbon atom or Carbon 1, which. Following a convention, C1 is considered as the one from the aldehyde to better understand the cyclization reaction and the differences on the chirality of the hydroxyl groups.

In one hand, carbohydrates will be called D- or L- according to the stereochemistry of the highest numbered chiral carbon. If this hydroxyl group is pointing to the right, it will be a "Dextro" carbohydrate and designated as D. If the hydroxyl group is pointing to the left, the sugar will be a "Levo" carbohydrate and designated as L. It is important to remember that most naturally occurring carbohydrates are of the D-configuration.

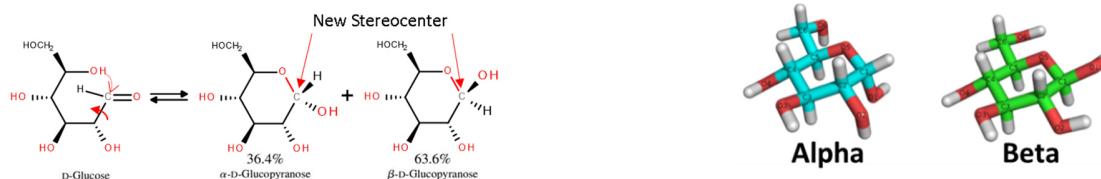
Significant differences between D- and L- enantiomers appears at cyclization, the chirality of the carbon 5 produces that, the bond between C1 and the oxygen from the hydroxyl group of C5 is located at different positions of the chiral center, giving the two possible and opposite chair conformations of each sugar molecule.



Cyclation of a D-Glucose and L-Glucose. Note that both molecules have exactly the same orientation as C1, C2, ... atoms are located exactly in the same positions.

In the other hand, the stereochemistry of the cyclization of a D- or L- enantiomers, can generate two different α or β stereomers, depending on the orientation of carbon 1 during cyclization. For example, for the cyclization of a D-glucose to a pyranose, where C1, becomes a stereocenter after cyclization, two stereomers of the pyranose are formed when it cyclizes, and they are called anomers.

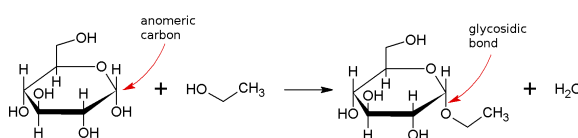
Then, depending on final orientation of the hydroxyl group of C1 after cyclization, we will designate these anomers as β - and α - carbohydrates. β -carbohydrates have a cis configuration between the OH group and the CH₂OH group. This means that the OH group and the CH₂OH group are on the same side of the ring. α -carbohydrates have a trans configuration between the OH group and the CH₂OH group. This means that the OH group and the CH₂OH group are on opposite sides of the ring.



Left, representation of the cyclization of α and β D-Glucopyranose, and right, representation of the α and β D-Glucose stereomers. Notice that, the orientation of the O1 oxygen atom marks the cis and trans configuration of the α and β isomers, the α isomer has the O1 atom in axial position and the β has it in equatorial.

THE GLYCOSIDIC BOND

Carbohydrates will be found isolated or forming polymers bonded through glycosidic bonds. Although more configurations can be found in nature, in general, glycosidic bond will be formed between the hydroxyl group of the C1 and any hydroxyl group from another carbon from the next bonding ring. Besides, there is a notation for naming the glycosidic bond in terms of the carbon atoms (C1, C2, ...) used for the glycosidic bond. For example, a glycosidic bond occurring between carbons C1 and C4 being C1 in β - configuration, will be designated as a " β -1,4" glycosidic bond.

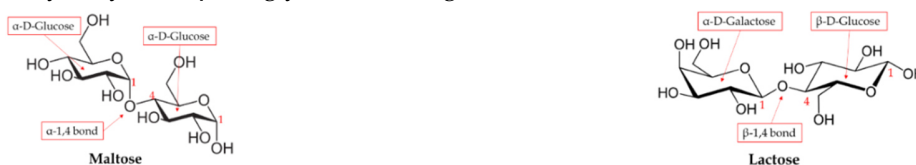


Formation of a glycosidic bond between monosaccharides. Image from Wikimedia Commons [7]

EXAMPLES: DISACCHARAIDES

Maltose is produced during digestion of starch and ultimately hydrolyzed (broken apart by water) into glucose to be used by the body. It is also produced in the manufacture of beer produced by malt producing the α -1,4-glycosidic linkage of maltose.

Lactose is major carbohydrate of mammalian milk, an individual who is lactose intolerant is deficient in the enzyme necessary to hydrolyze the β -1,4-glycosidic linkage in lactose.



Representation of Maltose (left) and Lactose (right) disaccharides remarking their α -1,4-glycosidic linkage and β -1,4-glycosidic linkage respectively. Image adapted from Wikimedia Commons [7].

DEGREE OF POLYMERIZATION

Polysaccharides may be classified according to their degree of polymerization, i.e. the number of monosaccharides linked together, and may be divided initially into three principal groups, namely sugars, oligosaccharides and polysaccharides [40].

Class(Degree of polymerization)	Subgroup	Components
Sugars (1–2)	Monosaccharides	Glucose, galactose, fructose, xylose
	Disaccharides	Sucrose, lactose, maltose, trehalose
	Polyols	Sorbitol, mannitol
Oligosaccharides (3–9)	Malto-oligosaccharides	Maltodextrins
	Other oligosaccharides	Raffinose, stachyose, fructo-oligosaccharides
Polysaccharides (>9)	Starch	Amylose, amylopectin, modified starches
	Non-starch polysaccharides	Cellulose, hemicellulose, pectins, hydrocolloids

EXAMPLES

POLYSACCHARIDES

A cellulose layer consists of numerous β -D-Glucose monomers connected by β -1,4-linkages. A strand of cellulose displays hydrogen bonds (dashed) within and between cellulose molecules and due to the β -1,4-linkages it displays a straight conformation contrary to the next example of amylose.

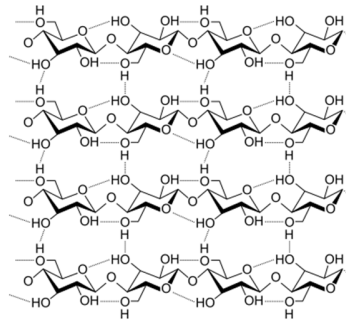


Figure showing a cellulose layer. Figure from Wikimedia commons [7].

Amylose consists of hundreds to about a thousand α -D-glucose monomers linked by α -1,4 glycosidic bonds. The straight chain forms a coil. (Figure A below). It is a major storage form of glucose in plants formed by a polymer of α -D-glucopyranose with α (1-4) glycosidic linkages.

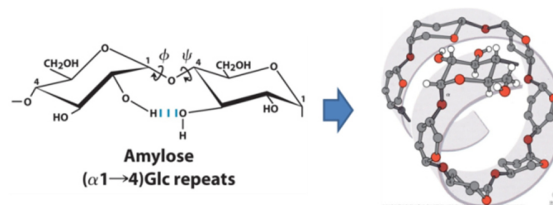
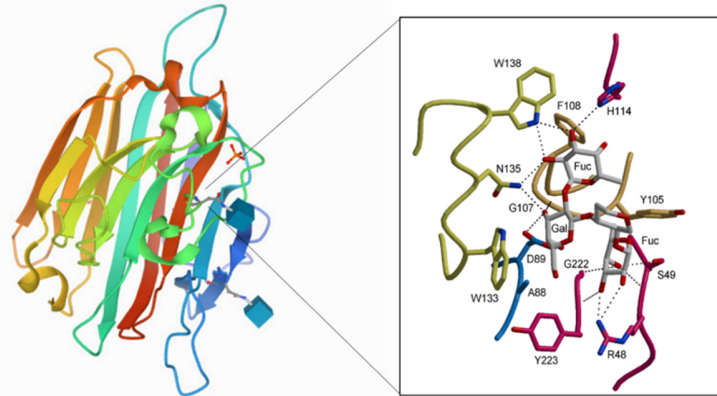


Figure showing an amylose strand with helical conformation due to internal hydrogen bonds. Adapted from Wikimedia commons [7].

LECTINS

Lectins are carbohydrate-binding proteins often found on the surface of cells, providing one mechanism of cell-cell recognition through binding of lectin on one cell to carbohydrate on another cell.



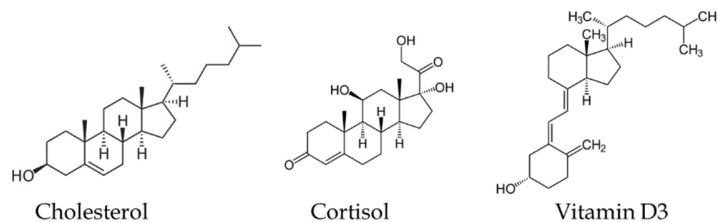
A disaccharide (shown in grey) bound in the binding site of a plant lectin. Image adapted from RCSB Protein Data Bank [1] and Wikimedia Commons [7].

LIPIDS AND PHOSPHOLIPIDS

Lipids are the group of biological macromolecules that have a major hydrocarbon component and are mostly nonpolar and hydrophobic. Functionally, lipids are important in cell membrane structure and in energy production. There are three main families of lipids, which are steroids, fatty acids and phospholipids.

STERIODS

Steroids are organic compounds having a core structure of four fused carbon rings. Cholesterol, many hormones as testosterone or estradiol, and vitamins are examples of steroids. Steroids have two principal biological functions. Certain steroids as cholesterol can have a structural function being important components in the cell membranes modifying the membrane fluidity. Other steroids are signaling molecules binding to hormone receptors.



Example of three steroid molecules. The cholesterol molecule (left) is located at the cell membrane having a structural role. The Cortisol (center) and Vitamin D3 (right) molecules are hormones involved in signaling. Image adapted from Wikimedia Commons [7].

LIPIDS AND FATTY ACIDS

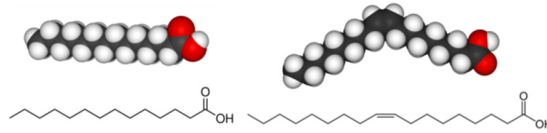
Fats can be found in form of single fatty acids, diglycerides or triglycerides that are glycerol linked to two or three fatty acid chains. Condensation reactions between glycerol hydroxyl groups and fatty acid carboxyl groups form ester linkages, joining the subunits.



Example of diglycerides (left) and triglycerides (right). R1, R2 and R2 denotes the possible substitutions of fatty acid chains with different lengths or unsaturation. Images from Wikimedia Commons [7].

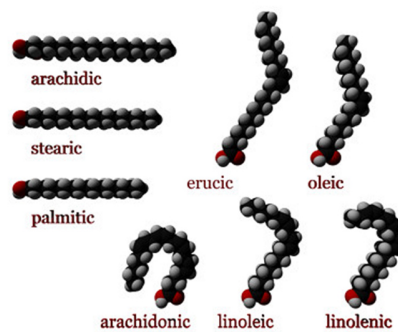
SATURATED AND UNSATURATED LIPIDS

Fats that have no double bonds in the hydrocarbon chain are “saturated” with hydrogens and are linear, *e.g.* animal fats. Fats that have a double bond are called unsaturated and have “kinks” in the hydrocarbon chain, *e.g.*, plant oils. Unsaturated fats can be either monounsaturated as shown or polyunsaturated (more than one double bond).



Left, saturated fatty acid chain with all single bonds between carbon atoms. Right, unsaturated fatty acid with one double bond, producing a bended conformation of the molecule. Adapted from Wikimedia Commons [7].

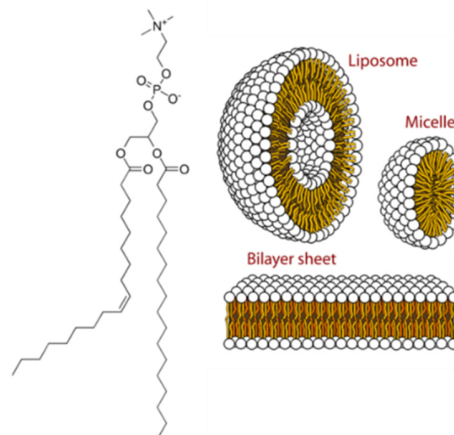
Note that, different fatty acids will receive a name depending on the number and position of the unsaturated carbon-carbon bonds.



Examples of fatty acids with unsaturations in different positions of the carbon chain. Image from Wikimedia Commons [7].

PHOSPHOLIPIDS

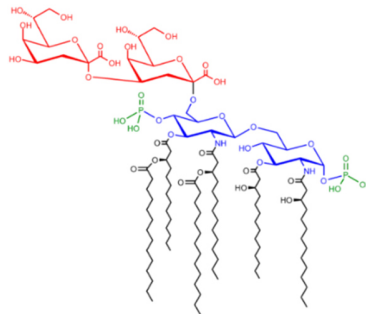
Phospholipids consist of a 3-carbon glycerol linked to a negatively charged phosphate group, and two fatty acids. Phospholipids are a major component of cell membranes due to their amphipathic nature, with a nonpolar or hydrophobic region, and a polar or hydrophilic region.



Self-organization of phospholipids: a spherical liposome, a micelle, and a lipid bilayer. Images from Wikimedia Commons [7].

SACCHAROLIPIDS

Saccharolipids are molecules composed by fatty acids chains bonded to the hydroxyl groups of a sugar backbone. Cell membrane bilayers integrate this type of molecules. For example, the acylated glucosamine precursors of the Lipid-A molecule, is one of the main component of the lipopolysaccharides in Gram-negative bacteria being components of the extracellular peptidoglycan wall.



Structure of the saccharolipid Kdo2-lipid A. Glucosamine residues in blue, Kdo residues in red, acyl chains in black and phosphate groups in green. Images from Wikimedia Commons [7].

CELL MEMBRANE

The plasma, cytoplasmic or simply the cell membrane is a set of biomolecules and separates the inside and the outside of cells, isolating them from the environment. Membranes are permeable to ions and some organic molecules, and it transports substances in and out the cell. It is composed of a phospholipid bilayer forming a big liposome containing the cytosol, nucleus and organelles from the cell. It has embedded membrane proteins, involved in many processes as cell adhesion, ion conductivity and cell signaling, saccharolipids, steroids, etc. The membrane is used as the attachment surface for several structures as the cytoskeleton.

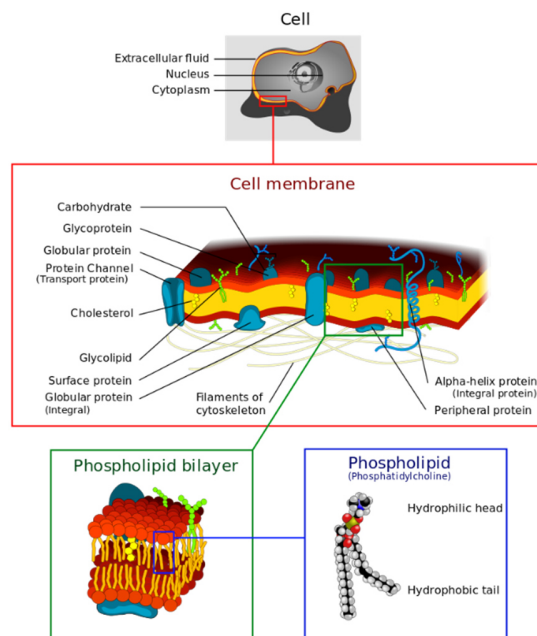
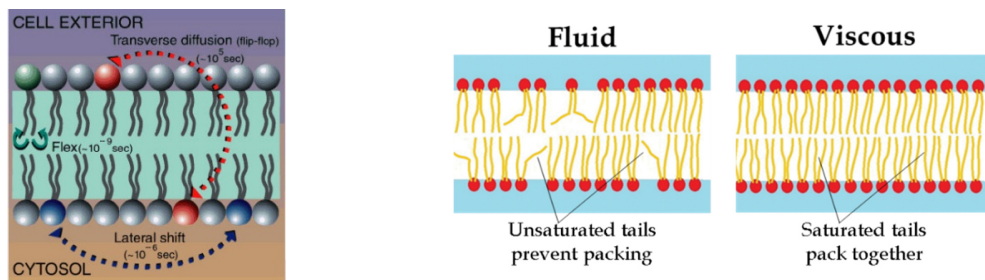


Illustration of the different levels of complexity of a Eukaryotic cell membrane. Images from Wikimedia Commons [7].

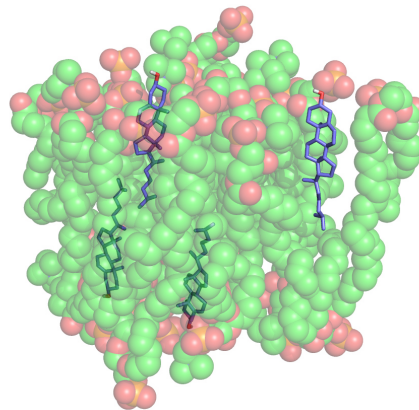
Phospholipids in the plasma membrane can move within the bilayer making the membranes act as fluids. Most of the lipids, and some proteins, drift laterally, and rarely the do molecule flip-flop transversely across the membrane. As temperatures cool, membranes switch from a fluid state to a solid state. The temperature at which a membrane

solidifies depends on the composition of lipids types. In addition, membranes rich in unsaturated fatty acids are more fluid than those rich in saturated fatty acids are.



Left, representation of the lateral shift and flip-flop movements of lipids that can occur in membrane Image from Wikimedia Commons [7]. Right, representation on how the unsaturated or saturated nature of the lipids present in the cell membrane, can affect to the fluidity. Image adapted from K. Nagy et al. [41] (CC BY 3.0)..

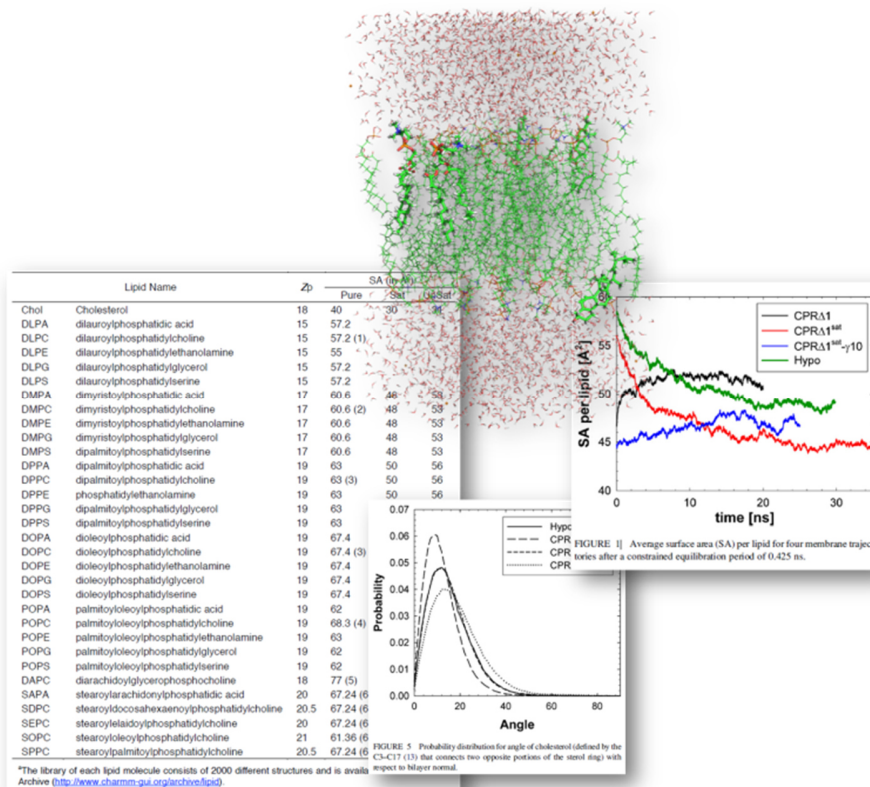
Membranes must be fluid to work properly being usually as fluid as salad oil. The steroid cholesterol has different effects on membrane fluidity at different temperatures. At warm temperatures such as 37°C, cholesterol restrains movement of phospholipids, and at cooler temperatures it maintains fluidity by preventing tight packing.



Insertion of cholesterol (represented in sticks and carbon atoms in blue) molecules in the cell membrane (represented in spheres with carbon from phospholipids in green).

Studying biological phenomena related to lipid membranes or membrane proteins in atomic detail has been of great interest to computational biophysicists, and there has been steady progress in molecular dynamics simulation studies of lipid membranes and membrane-associated proteins and peptides. MD simulations are particularly useful in such systems because they can provide the dynamics and energetics of membrane-associated proteins or peptides at the atomic level that is generally hard to obtain from experiments.

In model systems, the headgroup is usually phosphatidylcholine (PC) or phosphaditylethanolamine (PE), and tails generally lauroyl (L), myristoyl (M), oleoyl (O) or palmitoyl (P) chains. Combining these elements, we get lipids like DPPC (di-palmitoyl-phosphatidylcholine), DMPC (di-myristoyl-phoshatidylcholine), POPC (palmitoyl-oleoyl-phosphatidylcholine) and DLPE (di-lauroyl-phosphatidylethanolamine) [42].

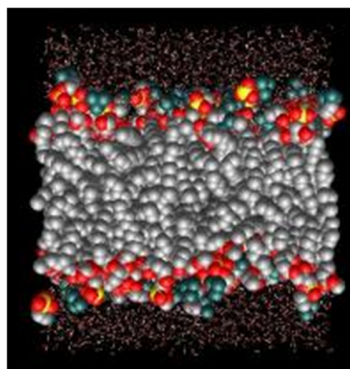


Representation of a cell membrane model, and some force field parameters from the glycam MMFF. Figure adapted from S.Jo et al. [42].

EXAMPLES: PHOSPHOLIPID BILAYER

From “Molecular dynamics simulations of fully hydrated DPPC with different macroscopic boundary conditions and parameters”, Tieleman *et al.* 1996 [43]:

“We compared molecular dynamics simulations of a bilayer of 128 fully hydrated phospholipid (DPPC) molecules, using different parameters and macroscopic boundary conditions. The same system was studied under constant pressure, constant volume, and constant surface tension boundary conditions, with two different sets of charges, the single point charge (SPC) and extended single point charge (SPC/E) water model and two different sets of Lennard-Jones parameters for the interaction between water and methyl/methylene.”

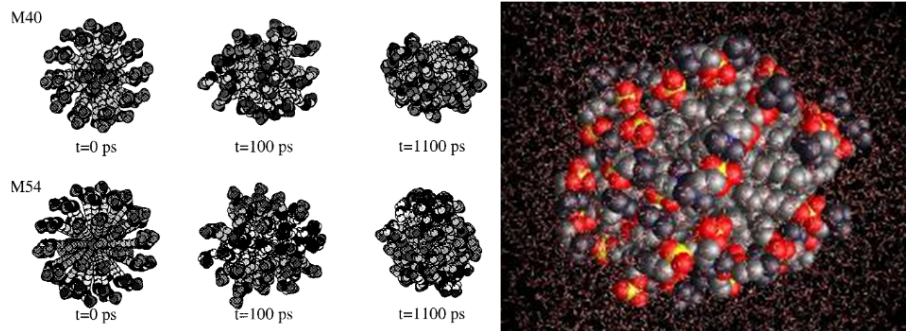


A bilayer consisting of 128 DPPC lipids and a few thousand water molecules (shown schematically). The structure is taken from ref. 1 and available for download.

EXAMPLE: MICELLES

Example from “Molecular dynamics simulations of dodecylphosphocholine micelles at three different aggregate sizes: micellar structure and lipid chain relaxation”, Tieleman *et al.* 2000 [44]:

“Simulation of 40 (M40), 54 (M54), and 65 (M65) dodecylphosphocholine (DPC) lipid micelles in water for up to 15 ns and analyzed the system.”

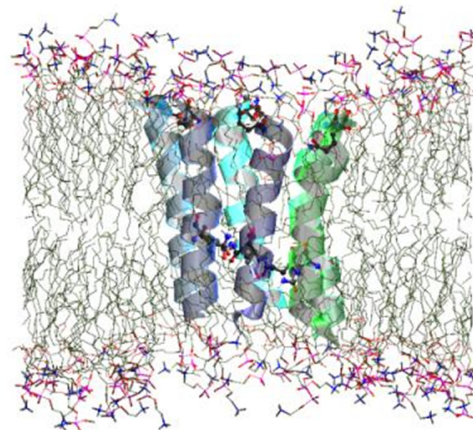


ADPC micelle consisting of 54 lipids and a few thousand water molecules. The structure available for download [43].

EXAMPLE: ALAMETHICIN

From “An alamethicin channel in a lipid bilayer: molecular dynamics simulations”, Tieleman *et al.* 1999 [45]:

“Results of 2-ns molecular dynamics (MD) simulations of a hexameric bundle of Alm helices in a 1-palmitoyl-2-oleoylphosphatidylcholine bilayer.”



Alamethicin helix bundle consisting of 6 helices in a POPC bilayer, taken from Tieleman *et al.* 1999 [45].

NUCLEIC ACIDS, DNA AND RNA

Deoxyribonucleic acid (DNA) and Ribonucleic acid (RNA) constitute the basic type of nucleic acids found in living organisms. They are large biopolymers made up of a linear array of monomers called nucleotides. Isolated nucleotides contain three components, a nucleic acid base, a ribose and a phosphate group.

The nucleic acid bases are derived molecules from pyrimidine and purine that are organic compounds formed by planar and aromatic heterocyclic nitrogenized bases. Pyrimidine is six member aromatic rings, and a pyrimidine ring fused to a five-member imidazole ring forms purines. There are three different pyrimidine derivatives, cytosine, thymine and uracil, represented by the letters C, T and U respectively, and two different purine derivatives, adenine and guanine, represented by A and G respectively, that differ in the nature and position of their substituents.

The ribose is a pentose carbohydrate in D configuration. As a convention, the carbon atoms of the ribose are primed numbers. Note that the D-ribose molecule, depending on whether carbon 2' has its natural hydroxyl group or not, will be named D-ribose or 2'-Deoxyribose.

Polymers of nucleotides formed by D-ribose or D-2'-deoxyribose will be called ribonucleotides or deoxyribonucleotides respectively, forming the Ribonucleic Acid (RNA) and the Deoxyribonucleic Acid (DNA). Deoxyribonucleotides and ribonucleotides can contain adenine, guanine and cytosine nucleic acids. However, uracil is only found in ribonucleotides and Thymine is only found in desoxiribonucleotides. Thymine and uracil are both of the same nature, but thymine has an extra methyl group.

When the nucleic acid base is found linked to the D-ribose or D-2'-deoxyribose sugar through an N-glycosidic linkage they are referred to as nucleosides. Purines bond to the C1' carbon of the sugar at their N9 atom, and pyrimidines bond to the C1' carbon at their N1 atom.

Finally, phosphate groups are added to the nucleoside to form nucleotides. We can find mono, di or triphosphate groups bonded to either C3 or C5 carbons of the ribose. This phosphate group plays a very important role since it is used to bond different nucleotides when they are found as a polymer (DNA or RNA) or is the source of energy when the hydrolyzing triphosphates to di-phosphates, for example hydrolyzing ATP to ADP.

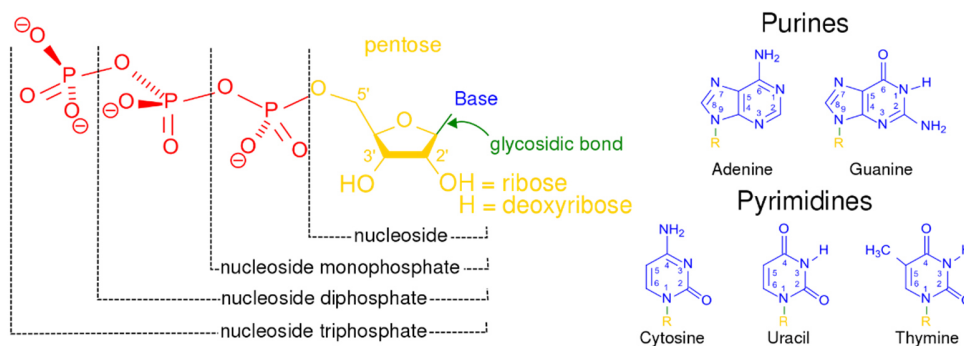
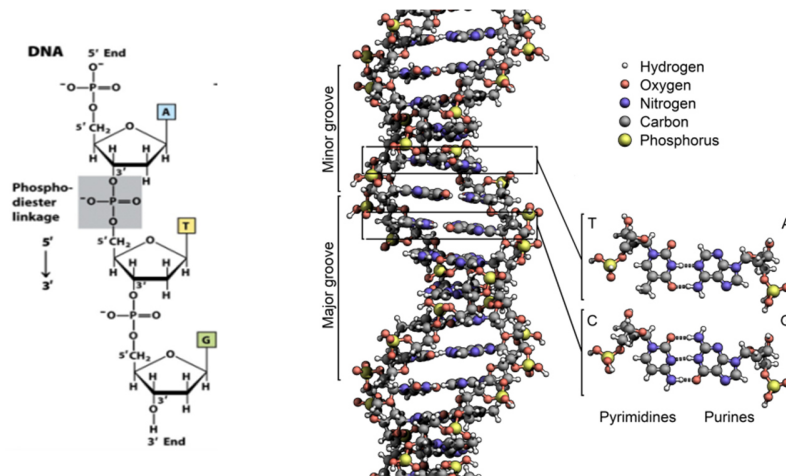


Figure showing the chemical representation of a nucleotide. Figure from Wikimedia Commons [7].

DNA

Phosphodiester bonds between consecutive nucleotides form the DNA molecule. This is a double stranded molecule forming an antiparallel filament having complementarity in sequence of amino acids and shape. Hydrogen bond forces link pyrimidines to complementary purines. Apart from these hydrogen bonds, there are stacking interactions between the pyrimidine and purine bases, making the characteristic helical conformation of the DNA with 10 base pairs per turn. This coiled conformation produces the characteristic minor and major grooves of the DNA. There are proteins that bind DNA to mediate in the transcription of DNA to RNA, or replication (copying DNA to DNA), and they make this interaction specifically with the major groove.

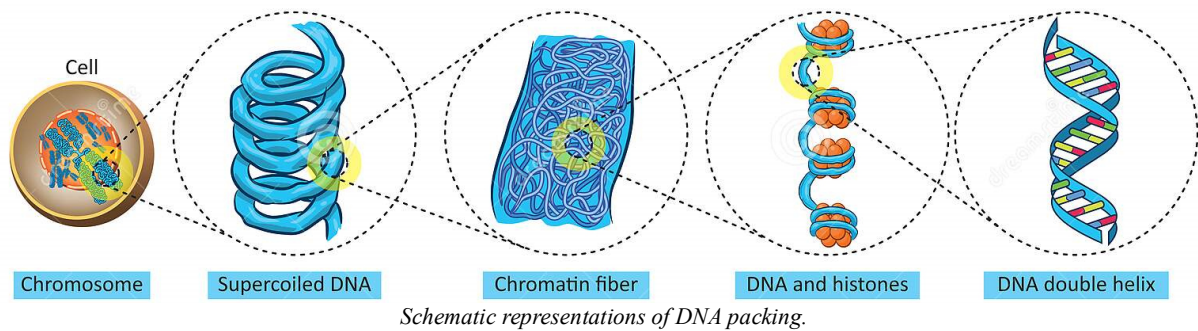
For the formation of the double strand, thymine interacts with adenine with two hydrogen bonds, and cytosine interacts with guanine with three hydrogen bonds. The hydrogen bond network established between the two strands makes DNA a very stable molecule.



Left, chemical representation of a single strand of DNA, remarking the 5'3' direction in terms of the phosphodiester linkage. Right, three-dimensional structure of DNA molecule with detail of the hydrogen bond interactions between pyrimidines and purines. Figure from Wikimedia Commons [7].

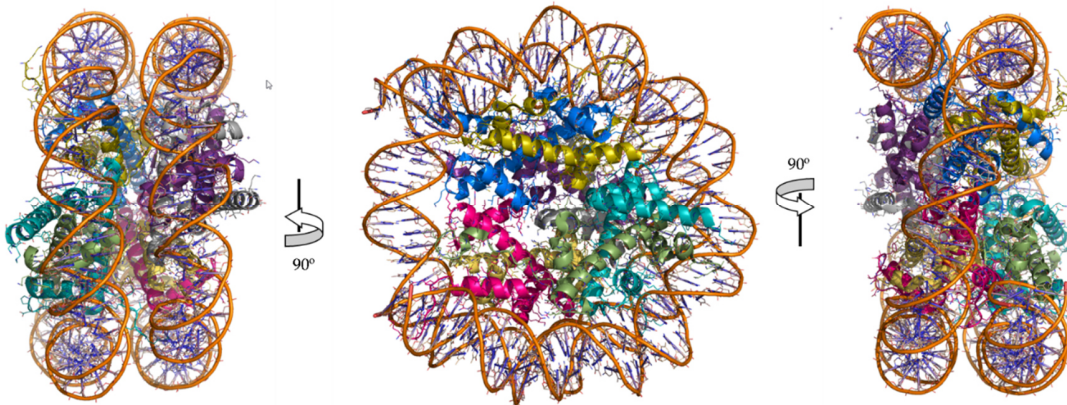
DNA PACKAGING: NUCLEOSOMES AND CHROMATIN

The DNA of a cell is packed into its microscopic nuclei with the help of histones, which help to organize DNA strand into the nucleosomes. Histones provide a structure where DNA can be wrapped around. Histones are positively-charged macromolecular complexes formed by eight proteins that strongly interact with the negatively-charged DNA. Nucleosomes are formed by a DNA strand rolled up 1.65 times around the histone. Furthermore, nucleosomes form a chromatin fiber of approximately 30nm, which assemble into loops averaging 300nm length. Finally this loops are compressed and folded to produce the chromatid fiber, ranging 250 nm wide, finally coiled into chromosomes[46].



EXAMPLE: NUCLEOSOME

Different experimental structures of nucleosome positioning have been already obtained. They display a sequence dependency which contributes to genomic regulation. In the next figure, the structures of nucleosome core particle are shown. The positioning is consistent with the central location of the minor groove inward regions contributing to minor groove bending, kinking and compression. The nucleosome center relates to a unique histone protein motif at this location, that enforces a sustained interaction with the narrow minor groove by hydrophobic interaction among others [47].



Nucleosome core particle, displaying ribbon traces for the DNA phosphodiester backbones (brown) and eight histone protein main chains. The views show the rotated DNA super helix. PDB-ID 3LEL.

THE GENOME

DNA molecules are finally packaged in the cell as structures called chromosomes. Bacteria have a single chromosome, but eukaryotes have multiple chromosomes, for example humans have 46 chromosomes (23 pairs) and about 3 billion nucleotide base pairs. All of an organism's chromosomes make up the genome. A single chromosome contains thousands of genes, each encoding a protein.



An image the organization of the 46 chromosomes making up the diploid genome of a human male. (The mitochondrial chromosome is not shown) Image from Wikimedia Commons [7].

THE GENES

Genes are fragments of DNA that contain the information for translating into an amino acid protein sequence. The information contained in a certain gene can be translated into computer language in the form of string of characters, using as an alphabet the initial letter of each nucleic acid base, A -> Adenine, C -> Cytosine, G -> Guanine, T -> Thymine, U -> Uracil. Commonly, genes start with the codon AUG that codify Methionine, the first amino acid in proteins.

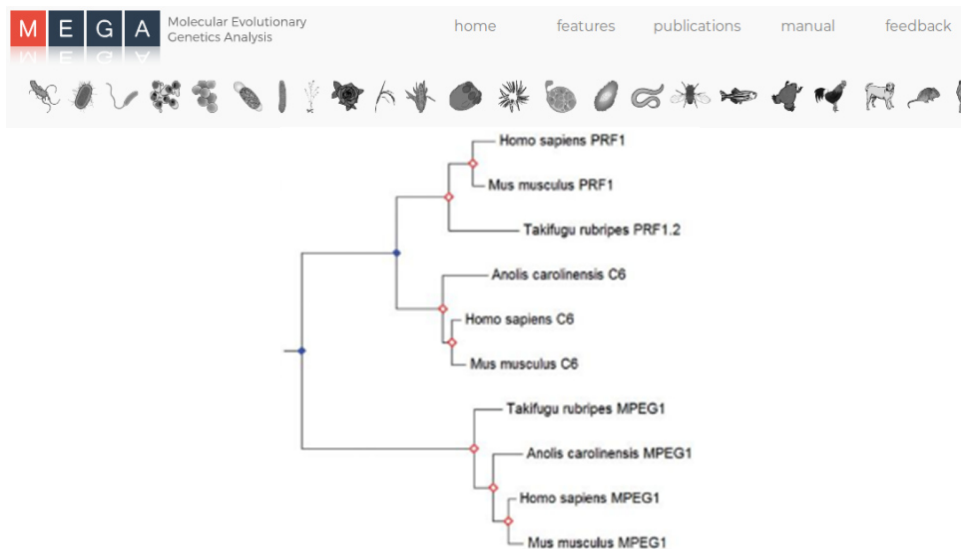
The most extended file format is the FASTA [48] file format for storing either the nucleotide or amino acid sequence of a gene or protein respectively. It contains an initial line started with ">" for annotations. Next, it follows by the nucleotide sequence.

```
>gi|109148525|ref|NM_000207.2| Homo sapiens insulin (INS), transcript variant 1, mRNA
AGCCCTCCAGGACAGGCTGCATCAGAAGAGGGCCATCAAGCAGATCACTGTCTTCTGCCATGGCCCTGTG
GATGCGCCTCCTGCCCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGAGCCTTTGTGAAC
CAACACCTGTGCGGCTCACACCTGGTGAAGCTCTCTACCTAGTGTGCGGGGAACGAGGCTTCTTCTACA
CACCCAAGACCCGCCGGGAGGCAGAGGACCTGCAGGTGGGGCAGGTGGAGCTGGGCGGGGGCCCTGGTGC
AGGCAGCCTGCAGCCCTGGCCCTGGAGGGTCCCTGCAGAAGCGTGGCATTGTGGAACAATGCTGTACC
AGCATCTGCTCCCTCTACCAGCTGGAGAACTACTGCAACTAGACGCAGCCCGCAGGCAGCCACACCCCG
CCGCCTCCTGCACCGAGAGAGATGGAATAAAGCCCTTGAACCAGCAAAA
```

DNA sequence in Fasta format.

GENETICS EVOLUTION

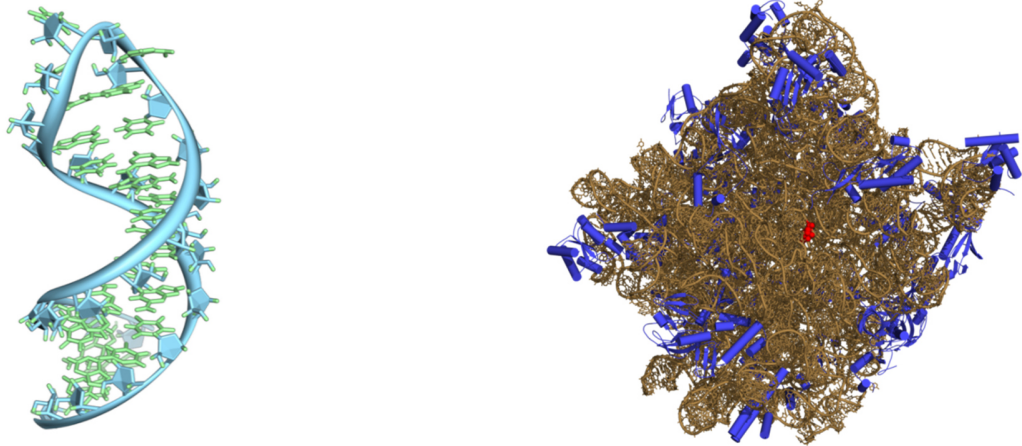
With its theoretical basis firmly established in molecular evolutionary and population genetics, the comparative DNA and protein sequence analysis plays a central role in reconstructing the evolutionary histories of species and multigene families, estimating rates of molecular evolution, and inferring the nature and extent of selective forces shaping the evolution of genes and genomes. These investigations have now expanded greatly due to the development of high-throughput sequencing techniques and novel statistical and computational methods. These methods require user-friendly computer programs that help to rationalize this big amount of data. One such effort has been to produce Molecular Evolutionary Genetics Analysis (MEGA) software [49].



Representation of the Molecular Evolutionary Genetics Analysis (MEGA) software [49].

RNA

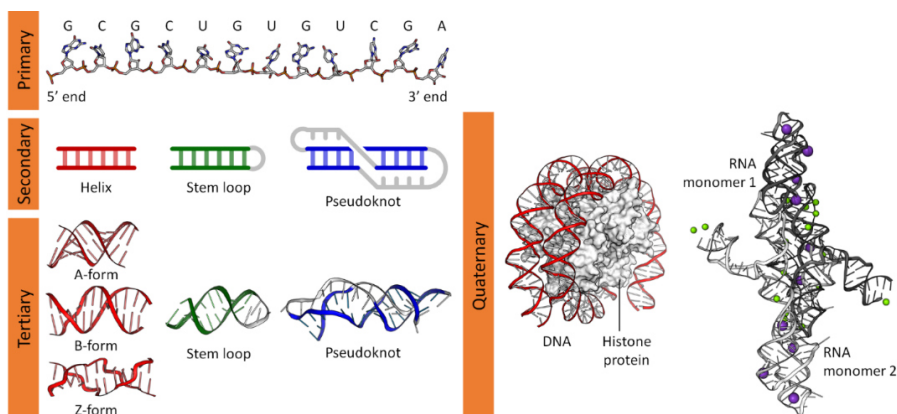
The transcription of a DNA molecule results in a single-stranded mRNA molecule that does not have a regular structure like DNA. Structures of RNA molecules are complex and unique making disordered interaction patterns. In RNA, we can also identify some secondary structure elements.



Left, Three-dimensional representation of the 50S ribosomal subunit. Ribosomal RNA is in ochre, proteins in blue. The active site is a small segment of rRNA, indicated in red. Right, Three-dimensional representation of the 50S ribosomal subunit. Ribosomal RNA is in ochre, proteins in blue. The active site is a small segment of rRNA, indicated in red. Image from Wikimedia Commons [7].

Guanine pairs with cytosine and uracil, and adenine will pair only with uracil. This will allow different regions of the RNA strand to interact with each other. Then we can identify some repetitive structural patterns identified as the secondary structure of the RNA. There are loop regions, helices where a several nucleotides of the same RNA strand are complementary, pseudoknots, internal loops, etc.

From these individual secondary structure motives, we can define the secondary structure of RNA molecule. Several bioinformatics tools allow you to predict the secondary structure of an RNA molecule, although in many cases it is not of great helps, since RNA is a complex molecule that adopts a very complex geometry and is involved in many biological processes still to unravel.



Left, secondary structure motives found in RNA, from loop, helix, harping loop and internal loop, and prediction of the secondary structure of a RNA sequence by a computer tool. Right, Summary of nucleic acid structure (primary, secondary, tertiary, and quaternary) using DNA helices and examples from the VS ribozyme and telomerase and nucleosome. Image from Wikimedia Commons [7]. (PDB-IDs: ADNA, 1BNA, 4OCB, 4R4V, 1YMO, 1EQZ)

PROTEINS

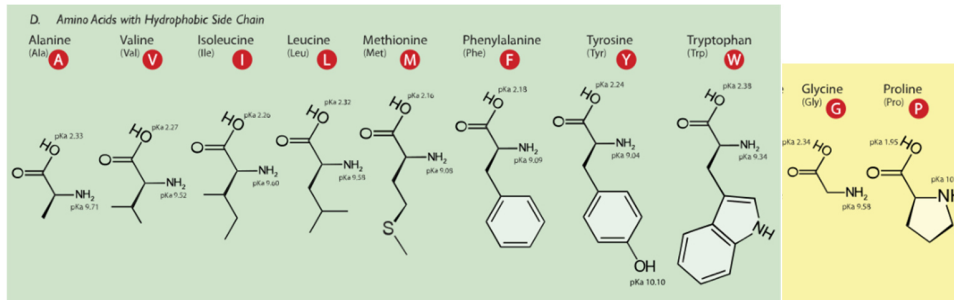
AMINO ACIDS

There are twenty different amino acids sharing a common main chain or backbone (NCCO), and varying side chain. Each amino acid has two different enantiomers. L-amino acids form proteins, although some bacterial amino acids are in D-form.



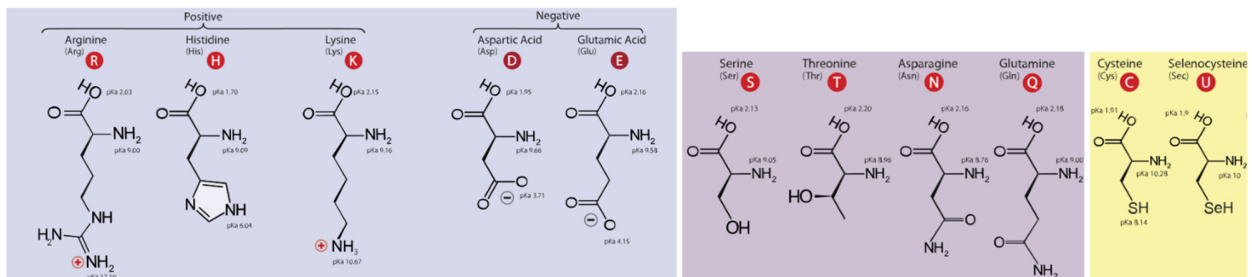
Left, schematic representation of an amino acid. Right, representation of D and L enantiomers. Image from Wikimedia Commons [7].

Amino acids are classified according to their chemical properties. Nonpolar amino acids, where side chains consist mainly of carbon-based groups. Leucine and isoleucine are structural isomers. Methionine has a Sulphur atom in its side chain. Sulphur has the same valence as oxygen. Phenylalanine and tryptophan have aromatic rings, which are flat due to the double bond network. Tryptophan is often defined as polar because of the NH group. In practice, however, it has more hydrophobic properties. Proline has its side chain a carbon group bound to the amino nitrogen to form a ring network making it more rigid than the rest of the amino acids.



Chemical representation of the non-polar amino acids. Green amino acids with hydrophobic side chain, yellow special cases. Image from Wikimedia Commons [7].

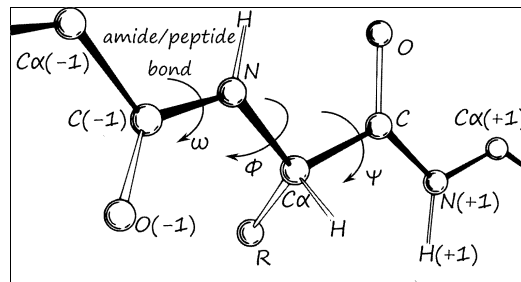
Polar amino acids, with side chain groups consisting of carbon, oxygen, and nitrogen atoms together, making the side chain more hydrophilic. Serine and threonine have hydroxyl functional groups. Cysteine has a thiol group (-SH) that is otherwise structurally similar to serine but not chemically similar. Charged amino acids are aspartic and glutamic with a carboxyl group, lysine with an amino group, arginine with a guanidine group, and histidine formed by an imidazole group, sometimes charged, most often classified as a polar amino acid.



Chemical representation of the polar and charged amino acids. Light purple amino acids with electrically charged side chains, dark purple amino acids with polar uncharged side chains, yellow special cases. Image from Wikimedia Commons [7].

PEPTIDE UNITS

A peptide is a set of covalently bonded amino acids where the covalent bond is called as peptide bond. Then, the ϕ /Phi, ψ /Psi and ω /Omega angles are the torsion angles resulting from bonding the main chain N- C_{α} , C_{α} -C and C-N atoms from consecutive amino acids respectively. ϕ /Phi and ψ /Psi can freely rotate allowing the formation of the secondary structure. Then, the ϕ angle is defined as the angle of right-handed rotation around N- C_{α} bond with values from -180 to 180 degrees, and ψ angle is defined as the angle of right-handed rotation around C_{α} -C bond with values from -180 to 180 degrees.

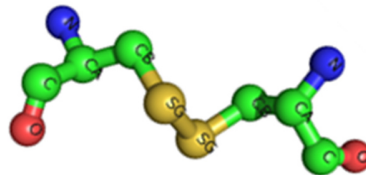


Representation of ϕ /Phi, ψ /Psi and ω /Omega angle. Image adapted from Wikimedia Commons [7].

The ω /Omega angle is formed by C-N atoms from a certain amino acid and its consecutive one, forming the amide bond. This bond cannot freely rotate as the other two, and only adopt *cis*- and *trans*- configuration always close to 180° for *trans*- peptides or 0° for *cis*- peptides ($\pm 30^\circ$ in extreme cases). *Cis*- peptides are energetically unfavorable because of steric clashes between the neighboring C_{α} atoms.

The only exception to this is the proline amino acid, where *cis*- peptide is just 4 times less favorable than *trans*- peptide, because there are some steric clashes in both *cis*- and *trans*- forms. Proline *cis-trans* isomerization is an important factor in protein folding, where special enzymes called prolylpeptidyl isomerases catalyze the transition from one form to another.

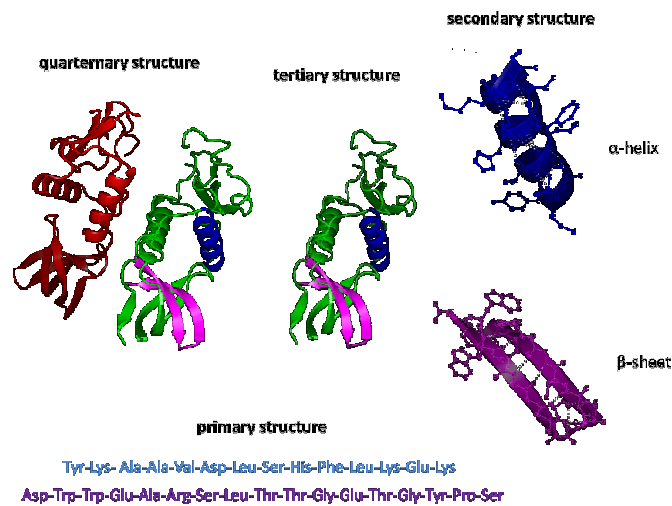
Disulfides bonds between cysteine amino acids are almost the only way to covalently link non-sequential amino acids. Formation of disulfide requires an oxidative environment. Therefore, disulfides are very rare in intracellular proteins but quite abundant in secretory proteins.



Representation of a disulfide bond.

THE STRUCTURE OF PROTEINS

The structure of a protein can be classified at primary, secondary, tertiary and quaternary level. The primary structure is defined by the amino acid sequence. The secondary structure is defined by the folding of α -helices or β -sheets. The tertiary structure is defined by the folding of the secondary structure motives in energetically stable subunits of protein chains. Finally, the quaternary structure involves the assembly of several tertiary structure subunits into protein complexes formed by different protein chains.

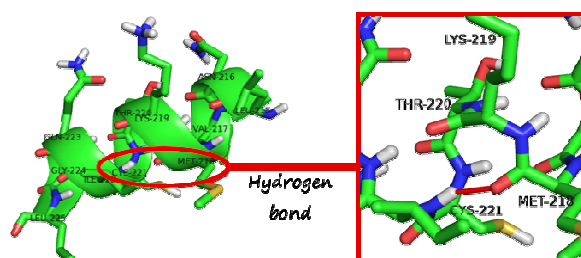


Representation of the different levels of classification of the structure of proteins. Image adapted from Wikimedia Commons [7].

The hydrophobic sidechains of protein have a tendency to cluster together in order to avoid unfavorable contacts with polar water molecules, then the interior of protein will be forming the hydrophobic core. Polar and charged amino acids are usually located on the surface of the protein being involved in hydrogen bond formation with other amino acids or with the solvent. They can also make hydrophobic contacts with their aliphatic carbon atoms. Anyway, these polar amino acids are rarely buried within the core of the protein.

The secondary structure of α -helices or a β -sheets is formed by efficient ways of hydrogen bond formation from the main chain atoms.

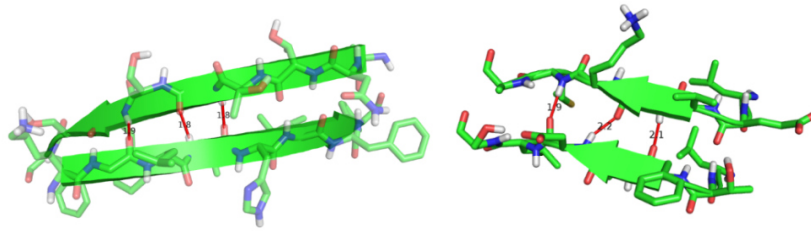
The α -helix has 3.6 residues per turn, and the hydrogen bonds are made between the hydrogen from the nitrogen and the carbonyl oxygen atoms from the amide bond of residues n and $n+4$. The most common location of a helix is along the surface of proteins, with one side of the helix facing the hydrophobic core and other side facing the solvent. Such a location results in a periodic pattern of alternating hydrophobic and polar residues. However, this pattern is not consistent enough for structure prediction, since small hydrophobic residues can face the solvent and some helices are completely buried or completely exposed.



Three-dimensional model of an α -helix. Intramolecular hydrogen bond in red.

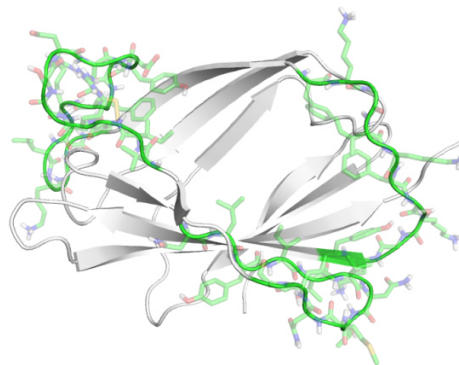
β -sheets are the other form of regular secondary structure in proteins. They consist of β -strands connected laterally by hydrogen bonds, between the hydrogen from the nitrogen and the carbonyl oxygen atoms from the amide at the backbone. They form a generally twisted sheet. A β -sheet is a section of the peptide chain typically consisting

from three to ten amino acids long, with backbone of the peptide chain placed in an extended conformation. Adjacent β -sheets can form hydrogen bonds in antiparallel, parallel arrangements.



Three-dimensional model of an antiparallel (left) and parallel (right) β -sheet. Intramolecular hydrogen bonds in red with their labeled distances in angstroms.

Loops are the remaining fragments of the chain connecting secondary structure elements. They are commonly located on the surface of protein. In general, main chain atoms nitrogen and carbonyl oxygen atoms do not make hydrogen bonds. Loops are rich in polar and charged residues and the length of loops can vary from two to more than 20 residues. Loops are very flexible, which makes them difficult to see in either X-ray or NMR studies of proteins. Within In homologous protein families, loop regions are less conserved than secondary structure elements. Insertions and deletions in homologous protein families occur almost exclusively in loop regions. Nevertheless, the sometimes they can be well conserved among families as they frequently participate in the ligand binding sites.

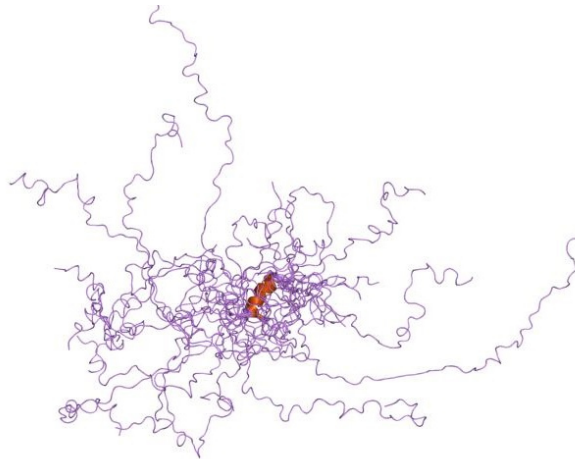


Structure of a protein, marking in green some of the loops found in the structure.

INTRINSICALLY DISORDERED PROTEINS

Intrinsically disordered proteins play an important role in cellular signaling, where the same peptide chain can perform adopt different conformations to make interactions with more than one protein. These disordered regions are often built from post-translational modifications and alternative splicing.

Experimental and computational analysis are usually combined to study, identify and characterize disordered regions. Intrinsically disordered proteins are composed by a sequence within the protein structure with biased amino acid composition with low content of hydrophobic amino acids. These sequences are unable to fold spontaneously into stable three-dimensional structures. They fluctuate rapidly over an ensemble of conformations that cover a continuum of conformations.



An ensemble of NMR structures of the Thylakoid soluble phosphoprotein TSP9, which shows a largely flexible protein chain. Image from Wikimedia Commons [7].

Most proteins contain intrinsically disordered regions over their peptide sequences. They combine them with stable globular domains. Intrinsically disordered proteins used in many different functions. They usually have a central role in signaling pathways, they participate in the assembly of molecular machines, in the assembly of microfilaments, in binding and transport of small molecules, etc. [50]

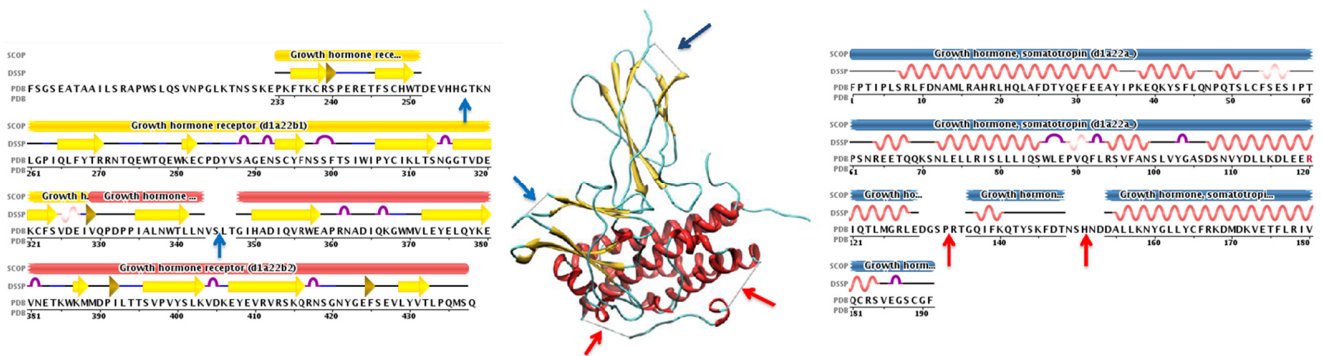
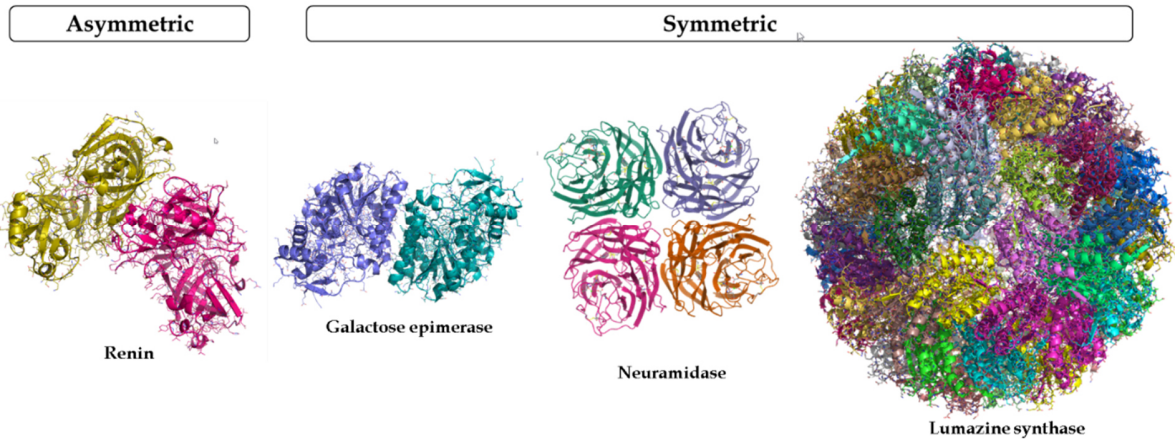


Figure showing missing electron densities in X-ray structure representing protein disorder (PDB-ID: 1A22). Compilation of screenshots from PDB database and molecule representation. Arrows point to missing residues. Image from Wikimedia Commons [7].

Prediction of disorder regions emerged as a field in structural biology leading to different bioinformatics tools as for example the DISOPRED server. It allows users to submit a protein sequence, and returns a probability estimate of each residue in the sequence being disordered. The results are sent in both plain text and graphical formats, and the server can also supply predictions of secondary structure motives to provide further structural information [51].

PROTEIN COMPLEXES

The assembly of individual monomeric proteins into quaternary structure, i.e. functional protein complexes, is crucial to almost all biological processes. Thousands of protein complex structures have been determined over the last decades improving our understanding of the biological rules that control quaternary structure organization into symmetric and asymmetric macromolecular structures [52].

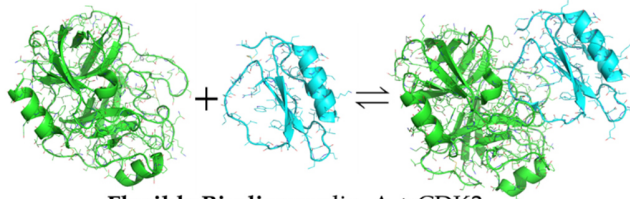


Examples from the different classes of asymmetric and symmetric molecular complexes. The PDB-IDs of crystal structures used in this figure are 1BIL, 1EK5, 1NNA and 3MK3[52].

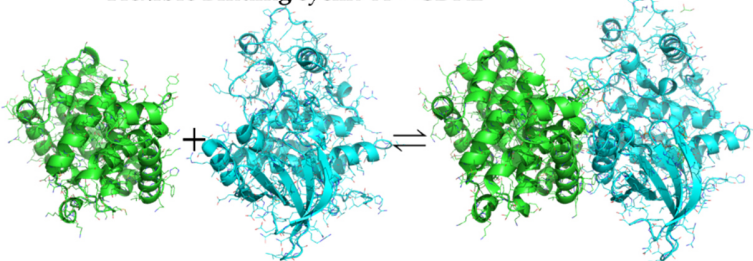
Furthermore, our conception of protein complexes has moved from a static representations of macromolecules, to a dynamic understanding of quaternary structure. Conformational changes include binding between molecules, multi-step ordered assembly pathways, and structural fluctuations occurring within fully assembled complexes.

Protein flexibility and conformational changes upon binding can be described. Rigid binding where no conformational changes occur in the protein subunits. Flexible binding where unbound state undergoes conformational changes upon binding. Finally, disordered binding between proteins, where intrinsically disordered proteins in its unbound state undergoes a major folding transition upon binding. [52].

Rigid Binding trypsin inhibitor + trypsin



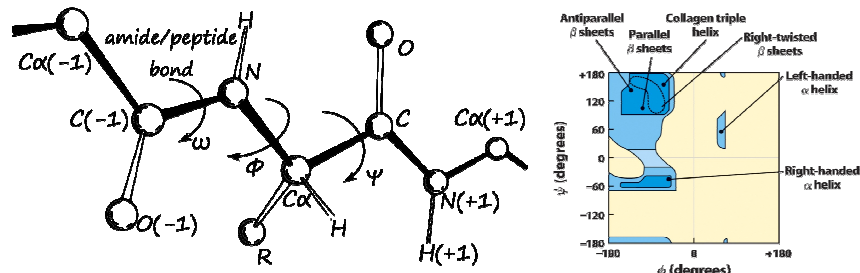
Flexible Binding cyclin A + CDK2



Representation of the rigid and flexible binding of trypsin inhibitor to trypsin and cyclin A to CDK2 respectively. The PDB-IDs of crystal structures used in this figure are 2A7H, 3RDY, 3RDZ, 2R3I, 1VIN and 2CCH [52].

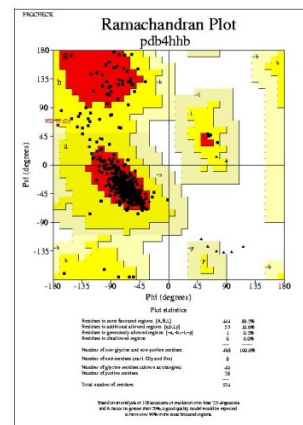
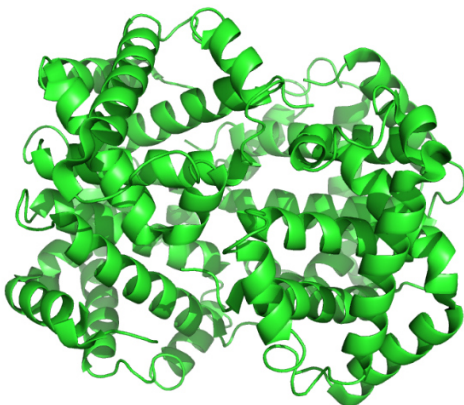
RAMACHANDRAN PLOT

A Ramachandran plot or a $[\varphi, \psi]$ plot [53], [54], is a way to visualize the most common combination of $[\varphi, \psi]$ angles observed in the secondary structure of proteins. The different left and right handed α -helices, parallel and antiparallel β -sheets, define certain regions in the plot where the combination of their φ and ψ dihedral angles fall. This approximation can be generalized for revising and validating the experimental structure or a model of a protein. Since certain combinations of $[\varphi, \psi]$ are allowed due to steric clashes of backbone atoms and $C\beta$ atoms, we should find the amino acids of our protein in clusters for specific regions except glycine.



Ramachandran plot, showing the different allowed regions for $[\varphi, \psi]$ angles depending on the secondary structure. Image adapted from Wikimedia Commons [7].

For example, the hemoglobin protein is almost formed by α -helices and its Ramachandran plot shows almost one only cluster of $[\varphi, \psi]$ angles.



Structure (left) and Ramachandran plot(right) of the hemoglobin (PDB-ID 4hhb). Image from RCSB PDB [1].

STORING THE INFORMATION OF PROTEINS

AMINO ACID SEQUENCE

The most extended file format for storing the sequence of amino acids of a protein is the FASTA [48] file format for storing either the nucleotide sequence of genes, as previously described, or amino acid sequences of protein. Equivalently to storing gene information, it contains an initial line started with ">" for annotations. Next, the amino acid sequence described in one letter code. Amino acids are represented by the following one letter codes:

A -> Alanine, B -> Aspartic acid (D) or Asparagine (N), C -> Cysteine, D -> Aspartic acid, E -> Glutamic acid, F -> Phenylalanine, G -> Glycine, H -> Histidine, I -> Isoleucine, J -> Leucine (L) or Isoleucine (I), K -> Lysine, L -> Leucine, M -> Methionine, N -> Asparagine, O -> Pyrrolysine, P -> Proline, Q -> Glutamine, R -> Arginine, S -> Serine, T -> Threonine, U -> Selenocysteine, V -> Valine, W -> Tryptophan, X -> any, Y -> Tyrosine, Z -> Glutamic acid (E) or Glutamine (Q).

PROTEIN 3D STRUCTURE

The information related with the 3D structure of macromolecules mainly consists the location of the coordinates of the atoms in 3D space. Along with the information about atoms locations, atom types and names, residue type and sequence number (where residue may refer to amino acids, nucleotide or other small molecules appearing in the 3D molecular structure), chain ID (where chain may refer to the different molecules or subunits contained in the 3D molecular structure), etc. These information may be available in several formats whereas the most common is the PDB.

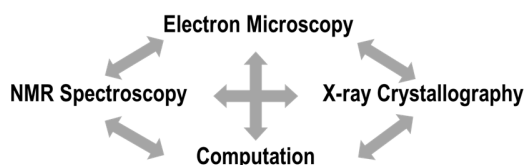
Apart from atomic information, PDB files include a header with different section describing information regarding experimental observations, bibliographic references, primary and secondary structure, etc. To explore the structures in the PDB file, it is helpful to use visualization programs as PyMol [55] or Chimera [25].

	Atom Nature	Atom number	Atom Name	Residue Name	Chain Id	Residue number	Coordinates			Additional Information		
ATOM	1067	NH1	ARG	A	141		-10.147	7.455	-6.079	1.00	23.24	N
ATOM	1068	NH2	ARG	A	141		-8.672	8.328	-4.506	1.00	33.34	N
ATOM	1069	OXT	ARG	A	141		-9.474	13.682	-9.742	1.00	31.52	O
TER	1070		ARG	A	141							
ATOM	1071	N	VAL	B	1		9.223	-20.614	1.365	1.00	46.08	N
ATOM	1072	CA	VAL	B	1		8.694	-20.026	-0.123	1.00	70.96	C
ATOM	1073	C	VAL	B	1		9.668	-21.068	-1.645	1.00	69.74	C
ATOM	1074	O	VAL	B	1		9.370	-22.612	-0.994	1.00	71.82	O
ATOM	1075	CB	VAL	B	1		9.283	-18.281	-0.381	1.00	59.18	C
ATOM	1076	CG1	VAL	B	1		7.449	-17.518	-0.791	1.00	57.89	C
...												
HETATM	1	C	ACE	C	0		50.950	33.338	48.783	1.00	42.49	C
HETATM	2	O	ACE	C	0		50.587	32.905	47.680	1.00	50.27	O
HETATM	3	CH3	ACE	C	0		50.361	34.676	49.132	1.00	49.32	C
...												
HETATM	2475	O	HOH	D	238		8.440	58.387	54.230	1.00	67.86	O
HETATM	2476	O	HOH	D	239		23.699	54.828	72.752	1.00	71.63	O
HETATM	2477	O	HOH	D	240		30.823	46.229	47.604	1.00	71.95	O

Representation of the atom information from a PDB file. Including Atom Nature (as standard atom from amino acid, nucleotide residues, or hetero atom belonging to other type of residues), Atom number; Atom Name (unique within each residue), Residue Name, Chain Id (identification character for the different macromolecule or small molecules), Residue number (within each chain), 3D coordinates, and some extra columns that can be used to include Additional Information.

EXPERIMENTAL STRUCTURES OF MACROMOLECULES

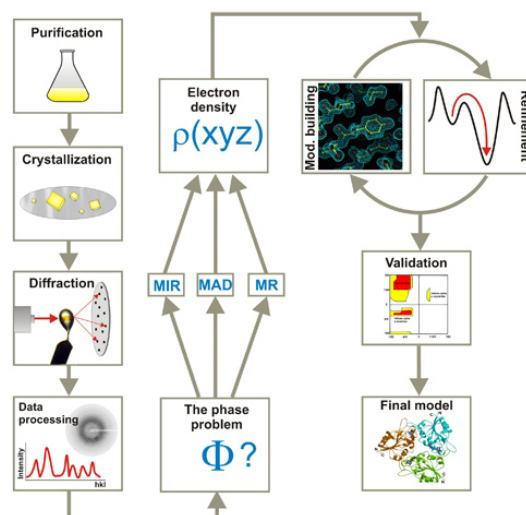
Structural Biology involves are several experimental and computational techniques for providing information of the three dimensional structure of the protein sequence. In order to obtain full understanding of the structural features of macromolecular complexes, and its relation with their biological function, a combination of these techniques is usually needed. Among them, we will review the three main technics that are closely related with molecular modelling. These are X-ray crystallography, nuclear magnetic resonance (NMR), and electron microscopy (EM). Each of them provides information at different levels, from coarse grain to atomic-level and are usually combined with computer calculations.



Schema of the different experimental and theoretical methods used to elucidate the structure of molecules, where arrows represent their complementarity.

X-RAY CRYSTALLOGRAPHY

X-ray crystallography of protein consists in, a series of steps leading to the location of the atoms in the 3D space. The purification of the protein sample and crystallization process, results in pure crystal then subjected to an intense beam of X-rays in order to obtain its characteristic diffraction pattern. The proteins in the crystal diffract the X-ray beam into one or another pattern of spots depending on the specific protein structure. The diffraction pattern is analyzed using complex methods that determine the phase of the X-ray wave in each spot, leading to the determination of the distribution of electrons in the macromolecule into a 3D map of the electron, i.e. electron density map. This map is then processed computationally to determine the location of each atom [56].

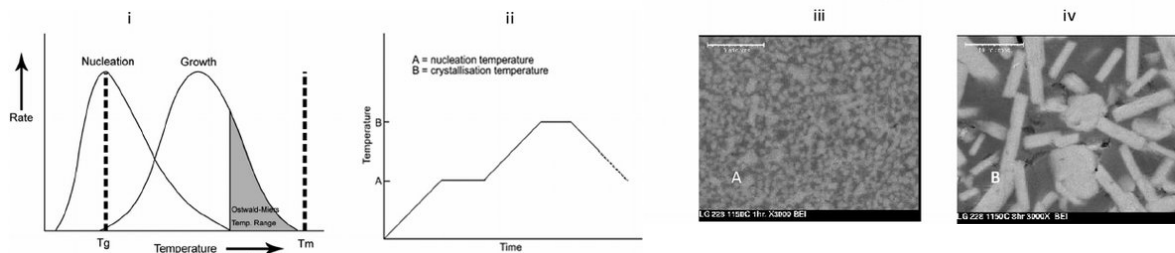


General diagram illustrating the process of resolution of molecular and crystal structures by X-ray diffraction. Image from M. Martínez-Ripoll [57] (CC BY 4.0).e.

Therefore, obtaining the structure of a protein by crystallography implies a series of steps that can be summarized as follows. First, molecular biology is used in order to over express the protein, normally using transfection of bacteria as an expression system and purification columns to obtain pure sample. Second, growing crystals of protein that diffract well that is a difficult step which can take from weeks to months until the process is properly optimized for the specific molecule under study. Third, experimental physics has to be applied to obtain a diffraction pattern of the crystal by applying an X-ray beam to the crystal and recoding the characteristic diffraction pattern. In order to use high precision X-ray beams usually experimentalists go to synchrotron facilities (particle accelerators), where thin X-ray beams appear tangent to the accelerated particle. Finally, computational models are applied to translate the diffraction pattern into a 3D electron density map with the use of fast furrier transforms, where protein amino acid atoms can be placed, with the prior knowledge of the protein amino acid sequence. Finally, refinement of the model by using force field based molecular energy minimization

CRYSTALLIZATION

The formation of protein crystals can happen when the purified sample is gradually brought into supersaturated state while reducing protein solubility by the addition of precipitants as polyethylene glycol and ammonium sulfate. This process is influenced by many factors as protein purity, pH, concentration of the protein, temperature, etc, that are difficult to predict in advance. Protein crystallization is set up with all these conditions plus using a vapor diffusion method as producing a “hanging” or “sitting” drop close to a reservoir in a sealing environment to allow the equilibration between the drop and the reservoir. The drop is checked periodically for the presence of crystals that may grow in hours, weeks or months.



(i) This schematically shows the rates of nucleation and crystal growth as a function of temperature. (ii) Two-step heat treatment is used for development of a crystal (iii) During the second hold at the crystal growth temperature nuclei grow into as seen in the SEM photomicrograph (iv). Reproduced from Ref. [58] with permission from the CNRS and The Royal Society of Chemistry.

The unit cell of the molecular structure forming the crystal is the smallest group of molecules forming a repeating pattern. The unit cell defines the symmetry and structure of the crystal lattice, which is built by the translation of the unit cell along its principal axes. There are different pattern of repetitions and they classified into space groups.

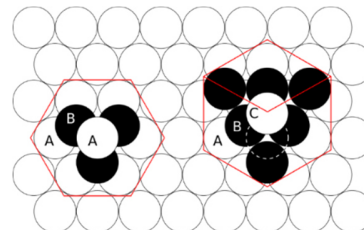
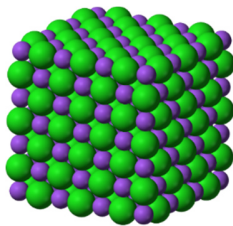
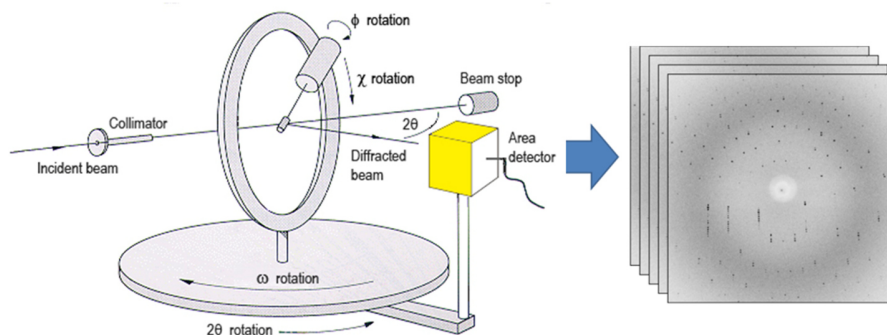


Figure showing, left, the crystallographic structure of sodium chloride, and two possible crystal lattice for the unit cell Figure adapted from Wikimedia Commons [7].

DATA COLLECTION

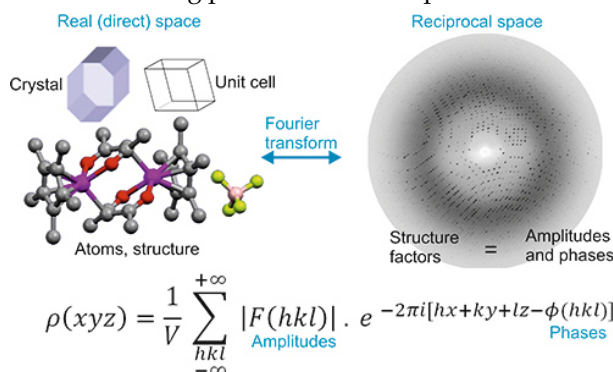
There are several experimental set-ups and protocols for obtaining the final model of the molecular structure contained in the crystal and represented by a single unit cell. All rely in the same physical principle, when a crystal is placed in an X-ray beam, it produces a diffraction pattern that is closely related with the spatial distribution of the electron of the atoms. Positioning the crystal in the X-ray beam should involve a system where the crystal can rotate for getting the diffraction patterns in the different perspectives. In a crystallographic experiment, the diffraction pattern forms a characteristic image of spots or reflections, where the intensity of these reflections is used to determine the distribution of electrons in the crystal.



Schema of Experimental set-up of X-ray. Image adapted from M. Martínez-Ripoll [57] (CC BY 4.0)..

Diffraction patterns reproduce the repetitions observed in the crystal at an atomic level, making the transformation of the spatial distribution electron density into a unique pattern. With the help of Fast Fourier Transform, a mathematical methodology able to decompose a signal into the frequencies and frequencies back into signals, whose absolute value represents the amount of that frequency present in the original function. In X-ray crystallography, the signal is the electron density, and the frequencies are represented by the diffraction patterns.

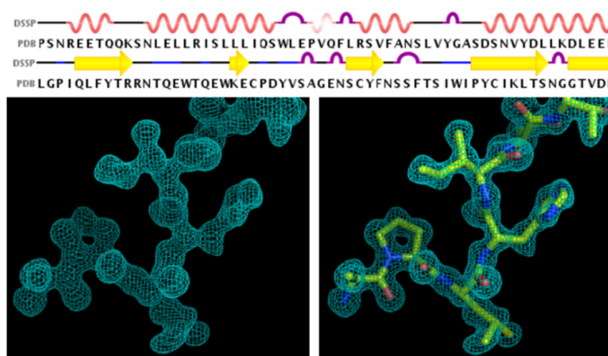
To create an electron density map we need two types of information, in one hand, the amplitude of X-rays in each reflection, and, in the other, the phase of X-rays in each reflection. This information defines the structure factor used to determine the electron density map. Measurement of the reflection intensities govern the amplitudes of the structure factors and the phases. There are several methods to estimate these values. The traditional approach is to add a few metal ions to the crystal, and matching the diffraction pattern with other crystals not containing such an atoms. The location of the heavy atoms can be found looking at these differences. Then, estimating phases based on their locations. Other common method to estimate phases is Molecular replacement, where previously-solved structures of the molecule are taken as a starting point to calculate phases.



Schema of Fast Fourier Transform (FFT), represented by the electron density function establishing a relation between the electron density maps and the spatial distribution of the electron density of the atoms in the crystal, direct and reciprocal spaces respectively. Image from M. Martínez-Ripoll [57] (CC BY 4.0)..

MOLECULAR STRUCTURE

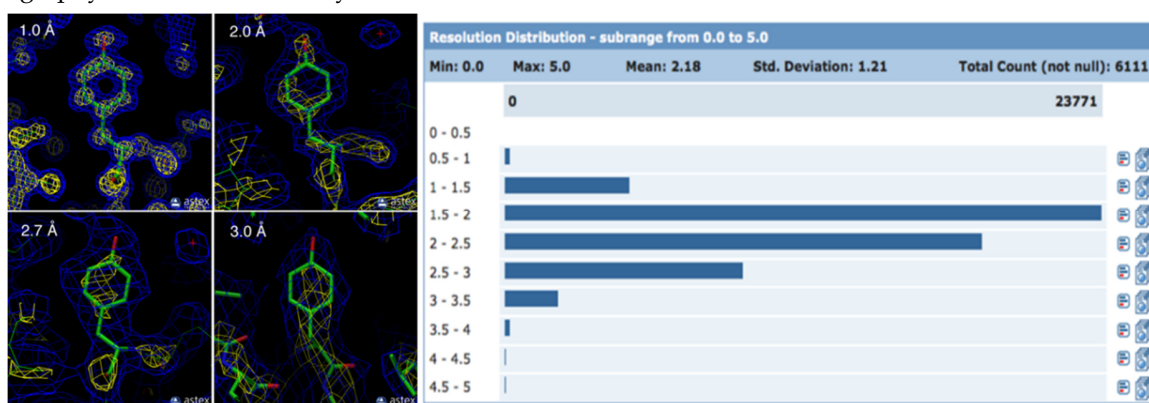
To determine a final model of molecular systems, as the ones accessible in databases as the RCSB Protein Data Bank [59], a combination of electron density map together with molecular mechanics energy minimization is required. This is largely automatic process by different computer programs, while it requires supervision during the whole process. Due to the wavelength of X-rays, hydrogen atoms are only resolved in the absolute highest resolution structures. So it is important taking into account that hydrogen atoms are missing in X-ray crystallography structures and normally added in posterior steps of molecular modelling.



Appearance of a zone of the electron density map of a protein crystal, before and after its interpretation in terms of a peptidic fragment. .
Image adapted from M. Martínez-Ripoll [57] under the terms of the (CC BY 4.0)..

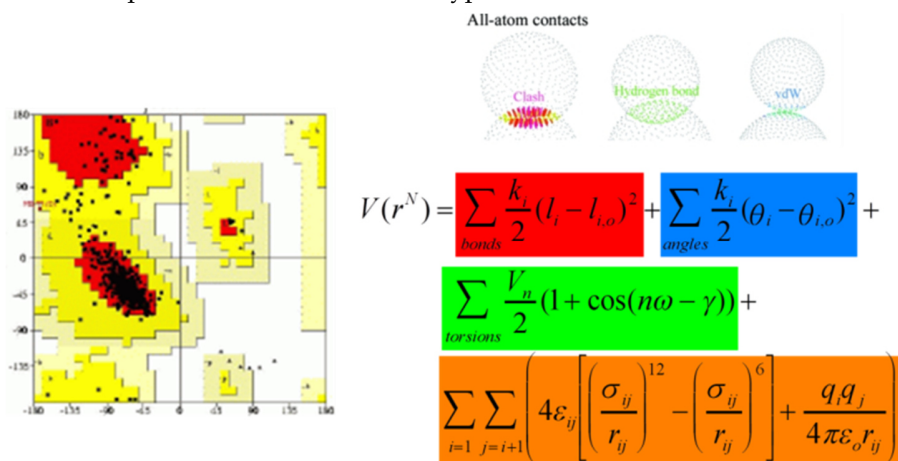
Depending on the quality on the crystal and the frequency of the X-ray used of the data collection, the electron density map will be of higher or lower resolution. High resolution means that the position of the atoms is located with more certainty than a lower resolution model. The resolution is measured in angstroms, and is a measure of the quality of the provided structure collected from the crystal. If all of the macromolecules in the crystal are sorted in an equal manner, then all of the molecules will scatter X-rays equivalently. In this case the diffraction pattern will display the details of crystal. If the molecules in the crystal do not have this same orientation the diffraction pattern will miss information of the detail.

Resolution is a measure of this level of detail from the diffraction pattern which will be mirrored when the electron density map is calculated. High-resolution structures have resolution values around 1 Å and are obtained from highly ordered crystals, being able to provide every atom location in the electron density map. Lower resolution structures can have resolution around of 3 Å or higher display the outlines of the protein chain. A good X-ray crystallography structure is normally considered when resolution lies above 2.5Å.



Left, representation of electron density maps at different resolutions. Right, bar-plot showing the frequency of atomic resolution of the electron density maps found in the Protein Data Bank. Image adapted from RCSB-PDB [1].

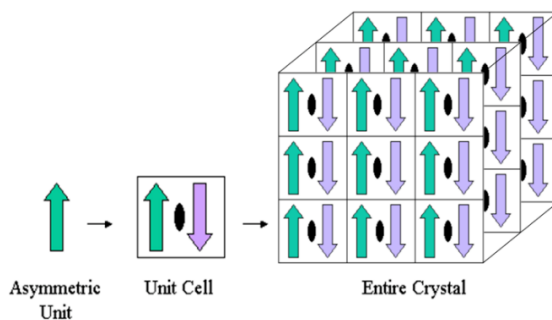
Finally, the Evaluation of models is performed by different approaches. Revision of the Ramachandran plot of the generated structure, energy evaluation using information provided by force fields, and parameters evaluation as for example the R-value that estimates the quality of the experimental structure obtained from X-ray crystallography [60]. The R-value measures how well a theoretical diffraction pattern generated from the 3D model matches the one observed experimentally. When the theoretical diffraction pattern is generated randomly, the R-value is about 0.63 whereas a perfect fit has a value of 0. Typical values in real scenarios are around 0.20.



Representation of the evaluation methods of final models obtained by crystallography, as inspection of Ramachandran plot and scoring by potential energy functions.

ASYMMETRIC UNIT CELL VS. BIOLOGICAL ASSEMBLY

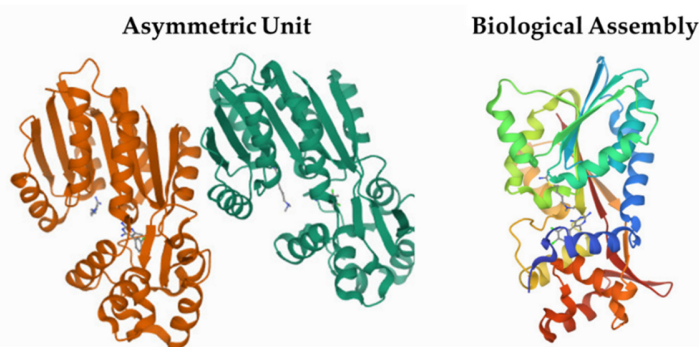
The information of experimental structures obtained by X-ray crystallography, as for example the Structure Summary provided by the RCSB PDB [59], usually reports images and coordinate files for the "biological assembly" and the "asymmetric unit". Sometimes both are the same, but you may find a difference between them in many cases. Normally the primary coordinate structure contains the crystal asymmetric unit that may or may not be the same as the biological assembly. Then, the asymmetric unit of the crystal is the minimum expression where the symmetry of the crystal is built by applying symmetry operations, i.e., rotations, translations and combinations of both. The biological assembly of the molecule is the functional form of the molecule. So a protein can be found as a dimer, trimer, tetramer, etc; in the asymmetric unit of the crystal, but be found as a monomer for its biological function, i.e. the biological assembly or the other way around. Application of symmetry operations to an asymmetric unit leads to one unit cell that makes up the entire crystal. Then, a crystal structure is formed by repetitions of unit cells that have as unique components the asymmetric unit parts.



Schematic representation of a molecule within the asymmetric unit, the unit cell and the entire crystal structure. Figure from PDB-101 [59].

One single biological assembly, a portion of a biological assembly or multiple biological assemblies, can form an asymmetric unit from a crystal. Moreover, the composition of such an asymmetric unit will depend on the orientation of the crystallized molecule together with the conformation adopted in the unit cell.

Depending on the experimental conditions, we may observe copies of the macromolecule or macromolecular complex from the crystal unit cell having identical conformations, or copies of the macromolecule, or macromolecular complex taking different conformations leading to unique positions over the crystal asymmetric unit. It is important to note that a biological assembly may be built from one single copy of the asymmetric unit, multiple copies of the asymmetric unit, or a portion of the asymmetric unit. The molecule can be seen as multimer within a crystal based on crystal packing, but may not form such a multimer in nature for performing its biological function.



Asymmetric unit and the biological assembly of the crystallographic structure of Histamine Methyltransferase Complexed with the Antifolate Drug Metoprine. PDB-ID: 2AOV

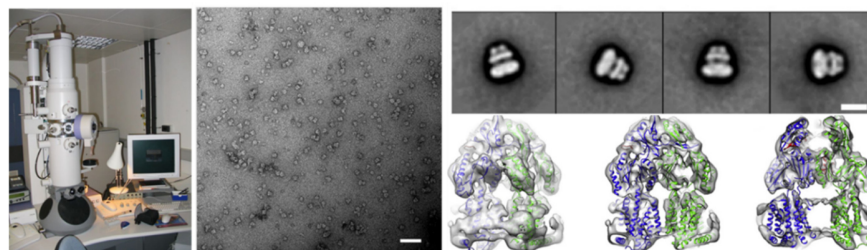
ELECTRON MICROSCOPY

Classic electron microscopy (EM) has been considered as one of the coarse grain methods, able to determine the structure of big molecular systems, from the structure of multimeric proteins to large molecular systems as molecular machines/motors, virus capsids or even at the tissue level [61].

Classic EM uses a beam of electrons to image the molecule and can be directly applicable to any protein or protein complex larger than 50kd, it provides a resolution up to 9Å, or even 7 Å for molecules with symmetry. The wavelength of an electron can reach very high frequencies in comparison with the wavelength of photons used in optical microscopies. The image resolution obtained by electron microscopies is much higher than any optical microscope, i.e. around 10.000.000x vs. 2000x respectively.

Normally, purified molecules compose the sample under study. Modern EM techniques comprehend cryo-electron microscopy (cryo-EM) for Transmission Electron Microscopy. Cryo-EM is based in preserving the biological sample close to its native state in an ice film at liquid nitrogen temperature or below. This thin layer of sample contain single molecules, where one single EM image contains several perspectives of the same molecule. There, we can apply computational algorithms to reconstruct a three dimensional representation of the surface of the molecules from a set of images from the different views. The reconstruction of the 3D structure is easier if the molecule has a certain degree of symmetry, such as in virus capsids. Cryo-EM needs small samples, in comparison with what needed for obtaining a pure crystal, holding for flexible or heterogeneous molecules, treated later during the image processing.

A simplification of a classification algorithm of the image containing several perspectives of the same molecule could resumed as follows. First, the acquisition of different EM images of the same sample. Second, identification of individual molecules. Third, classification and alignment of all the images of individual molecules into the different perspectives of the molecule. Forth, refinement of the image from each perspective by averaging the image from each set of images previously classified. Fifth, reconstruction of the 3D structure from the different refined perspectives.

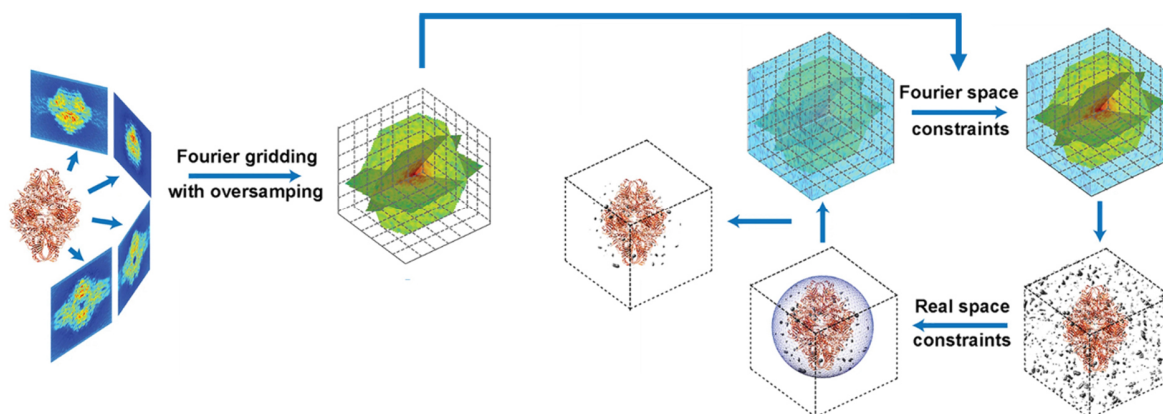


Electron microscopy image acquisition showing, from left to right, an electron microscope, electron microscopy images, classification of particle views in the image and the final 3D model of the protein. Image adapted from M. Parmar et al. [62] and Wikimedia Commons [7] (CC BY-NC-ND 4.0)..

Recent developments in cryo-electron microscopy have led to important advantages in this field, which have provided new generation detectors able to obtain resolutions at atomic level equivalent to the results obtained by X-ray crystallography or NMR [63], [64].

These detectors identify electrons directly instead translating them first into photons later translated again into photoelectrons as old detectors used to do, but also combining the advantages of charge-coupled device cameras and film [65]. These detectors were possible using a pixel sensor technology similar to camera chips but resistant to electron radiation. Detectors have up to 1.6 million pixels whereas they are bigger than any other standard detector to prevent electrons from activating more than one, and also they are integrated in an extremely thin layer avoiding blurring the image acquisition and allowing a very fast image acquisition compensating displacements produced from the electron beam forces.

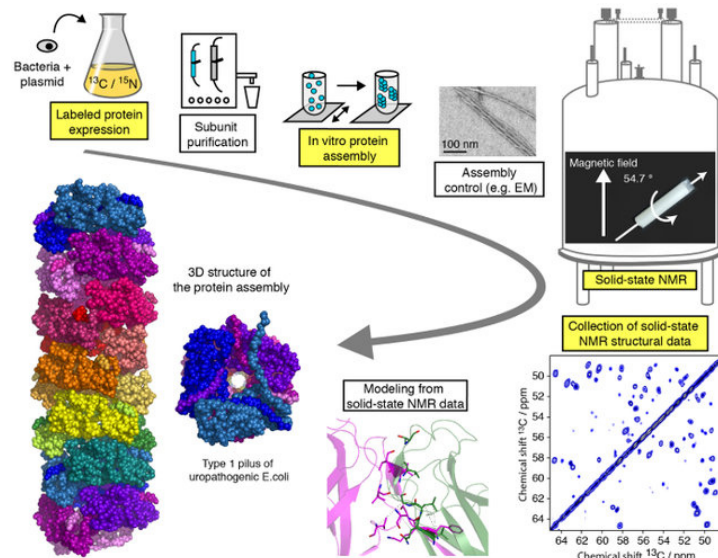
These high-resolution images obtained by cryo-EM can be used to generate accurate electric potential maps in the 3D space for the frozen macromolecules, equivalent to the electron density maps derived from X-ray crystallography. The electric potential maps are generated using the projection-slice theorem that states that the two dimensional projection of a three dimensional object in real space is equivalent to taking a central slice in two dimensions out of the three dimensional Fourier transform of that object. Obtaining central slices in several orientations crossing the center of the molecule, allows reconstructing the volumes where atoms from the macromolecule are placed leading to the electric potential map. This map is then used to fit the sequence or atoms from the macromolecule under study similarly as it is done in X-ray crystallography.



Representation of a tilt series of 2D projections from protein in the real space transformed into Fourier slices, used to calculate a small fraction of points on a 3D Cartesian grid leading to an electric potential map. Image from Alan Pryor Jr. et al[66] (CC BY 3.0)..

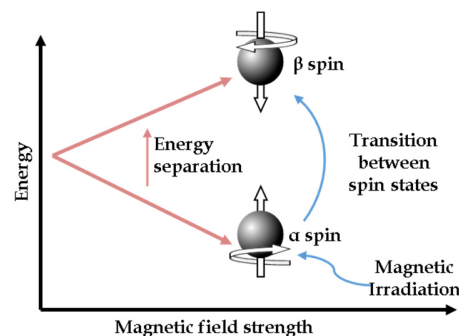
PROTEIN NMR

Nuclear Magnetic Resonance spectroscopy of proteins (NMR) is an experimental technique used in structural biology where we obtain information about the structure and dynamics of small molecules, proteins, nucleic acids or molecular complexes. Their measurement provides a map of how the atoms are chemically linked, how close they are in space, and how rapidly they move with respect to each other. This is done by analyzing the sample of a solution of highly concentrated purified molecules.



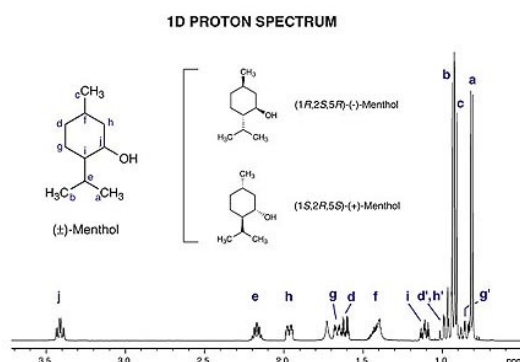
Representation, of the workflow for the data analysis Nuclear Magnetic Resonance (NMR). Labeled protein production, subunit purification, subunit assembly, control of assembly formation, NMR experiments, experiment analysis, extraction of distance restraints, and structure modelling. Image from A. Loquet et al. [67] (CC BY 4.0)..

This technique uses the fact that certain atomic nuclei are magnetic, giving to this, limited number of isotopes a property called spin is used. NMR can analyze the different isotopes, but the simplest example is the proton of the hydrogen nucleus (^1H). The spinning of the proton generates a magnetic moment. If we applied a magnetic field to the sample the spinning of the proton has two possible spin states, i.e. orientations of the magnetic moment usually called α and β), where one of the orientations is more populated than the other for a certain wavelength because it is aligned with the field (α state). Then if we provide an electromagnetic radiation pulse at a certain wavelength producing the energy needed to change the spin of the proton to the excited state (β state), we would obtain the resonance energy of the proton. Then a resonance spectrum of a molecule is obtained by the variation of the magnetic field a constant frequency of electromagnetic radiation, or by keeping the magnetic field constant and varying the electromagnetic radiation [68].



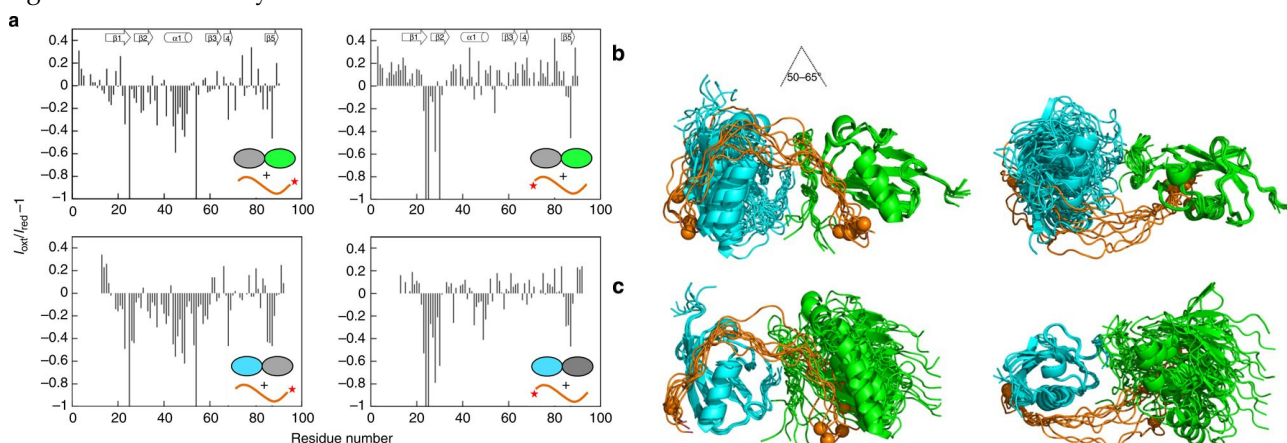
Left, representation of the α and β states of a isotope. Right, table showing the spin number, gyromagnetic ratio of the common isotopes found in biochemistry.

These properties can be used to examine the chemical surroundings of the hydrogen nucleus. The flow of electrons around a magnetic nucleus generates a small local magnetic field that opposes the applied field. The degree of such shielding depends on the surrounding electron density. Consequently, nuclei in different environments will change states, or resonate, at slightly different field strengths or radiation frequencies. The different frequencies, termed chemical shifts, are expressed in fractional units δ (parts per million) relative to the shifts of a standard compound. For example, a $-CH_3$ proton typically exhibits a chemical shift (δ) of 1 parts per million (ppm), compared with a chemical shift of 7 ppm for an aromatic proton. The chemical shifts of most protons in protein molecules fall between 0 and 9 ppm. It is possible to resolve most protons in many proteins by using this technique of one-dimensional NMR. With this information, we can rebuild changes and assign them to a particular chemical group under different conditions, such as the conformational change of a protein [69].



Example 1H NMR spectrum (1-dimensional) of a mixture of menthol enantiomers plotted as signal intensity (vertical axis) vs. chemical shift (in ppm on is of the horizontal axis). Signals from spectrum have been assigned hydrogen atom groups (a through j) from the structure shown at upper left. Image from Wikimedia Commons [7](CC BY 4.0)..

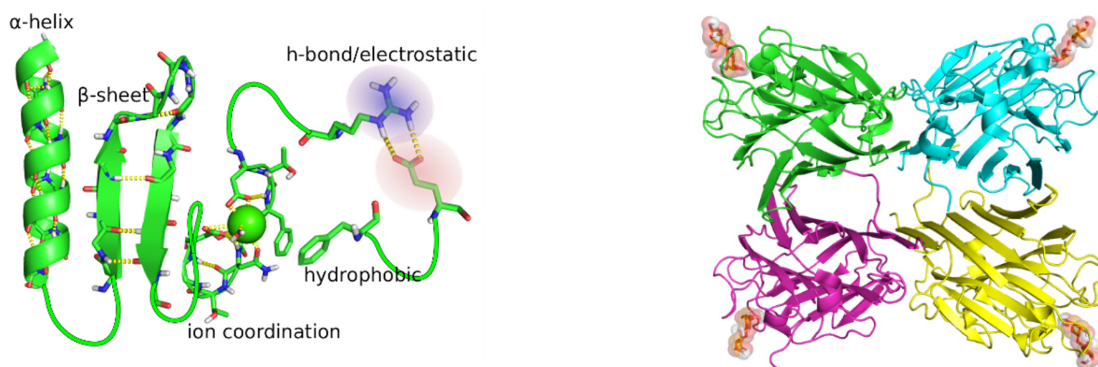
Identifying resonance pics of larger molecules is a challenging task, but it can be useful for detecting differences between conformations of proteins, mutations of the same protein, etc. We can identify whether hydrophobic amino acids could be interacting with other atoms or not. In addition to structural information, nuclear magnetic resonance can yield to insights on the dynamics of various parts of the protein. All this makes NMR remarkably different to X-ray crystallography or electron microscopy, by providing a dynamic view of the protein giving insights on the flexibility of the macromolecular structures.



(a) Plot of intensity difference between spectra for SUMO-2 dimers in complex with spin-labelled SIM2,3 peptide compared with the spectra in which MTSL is reduced by ascorbic acid. Data for complexes in which the distal subunit SUMO-2 $\Delta N11$ is labelled are presented on the bottom row, whereas those for labelled proximal subunit SUMO-2 ΔGG are shown on the top. Data for complexes with spin-labelled SIM2,3 with MTSL positioned at C terminus are shown on the left and with MTSL at the N terminus on the right. (b) The NMR ensemble of the lowest energy 10 NMR structures for the complex of SUMO-2 dimer with SIM2,3 in which the proximal subunits (SUMO-2 ΔGG) are superposed. (c) The NMR ensemble of the lowest energy 10 structures for the complex of SUMO-2 dimer with SIM2,3 in which the distal subunits (SUMO-2 $\Delta N11$) are superposed. Image from Y. Xu et al. [70] under the terms of the (CC BY 4.0)..

MOLECULAR INTERACTIONS

Molecular interactions are the physical principles that govern the structure of macromolecules and its stability. We find interactions between atoms within the protein chain, and interactions between the protein and the solvent, ions, metals ligands. They can be of an attractive or repulsive nature, favoring different physicochemical events and states as the aggregation of hydrophobic amino acids in the core of the protein, or the hydrogen bonds formed during the protein folding process, or the binding between monomers of small molecules of a macromolecular complex [71].



Left, schema of some of molecular interactions found in the structure of proteins. Right, crystal structure of a recombinant *Vatairea macrocarpa* seed lectin complexed with lactose (PDB-ID: 4WV8).

BONDED INTERACTIONS

Bonded interactions are mainly based by covalent bond, leading to the internal energy previously defined in Molecular Mechanics Force Field. A covalent bond is a chemical bond that involves the sharing of electron pairs between atoms, based in a stable balance of attractive and repulsive forces between atoms when they share these electrons.

Apart from covalent bonds, another type of important bonded interactions helping to maintaining the tertiary structure of proteins are disulfide bond. These interactions are derived by the coupling of two thiol groups from cysteine amino acids. They play an important role in the folding and stability of some proteins, usually proteins secreted to the extracellular medium. Since most cellular compartments are reducing environments, in general, disulfide bonds are unstable in the cytosol, with some exceptions as noted below, unless a sulfhydryl oxidase is present



Left, 3D representation of the structure of an isolated leucine amino acid. Right, representation of a disulfide bond between two cysteine residues.

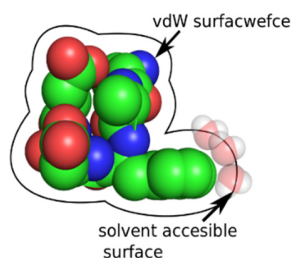
NON-BONDED INTERACTIONS

THE VAN DER WAALS (VDW) INTERACTION

As defined in previous section, when two molecules come into close contact vdW interactions appear accounting for repulsion, which act at small distances between the partners due to the overlap of electron clouds, or due to correlations between electrons in different atoms (London's dispersion forces).

Then, when defining vdW interactions between molecules we are defining the surface of the molecule based on the spheres defined by the van der Waals radii of each individual atom. Since the interatomic distances between bonded atoms are less than the sum of the atomic van der Waals radii, the volume inside the vdW surface is smaller than the sum of the vdW volumes of the atoms. Furthermore, molecules and proteins can be viewed as pictorial overlap of the spherical van der Waals surfaces of the individual atoms.

The solvent-accessible surface area (SASA) of a molecule is calculated using the "rolling ball" algorithm [72], which uses a sphere representing the solvent of a particular radius, to "probe" the surface of the molecule. To compute that SASA, a mesh of points is defined, then the number points that are accessible to a solvent molecule describe the surface area. The points at a certain distance of a molecule are estimated by radius beyond the van der Waals radius of a water molecule, that is similar to 'rolling a ball' along the surface. Then, all points are revised checking if each mesh point is buried or accessible. Finally, the portion of surface area each point represents to calculate the SASA multiplies the number of points accessible.



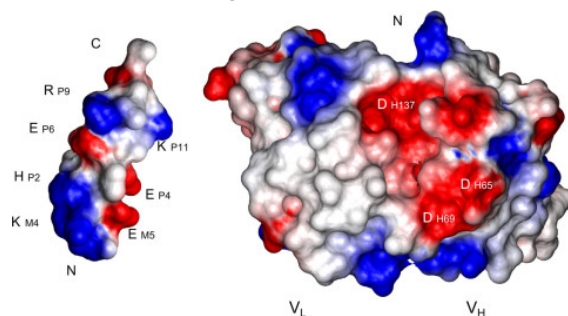
Representation of the vdW surface and solvent accessible surface.

THE ELECTROSTATIC INTERACTION.

As commented in previous sections, electrostatic interactions take place in a great number of binding events as charge-charge interactions, hydrogen bonding, p-p stacking, hydrophobic interactions, solvation among others.

ELECTROSTATIC SURFACE OF A PROTEIN

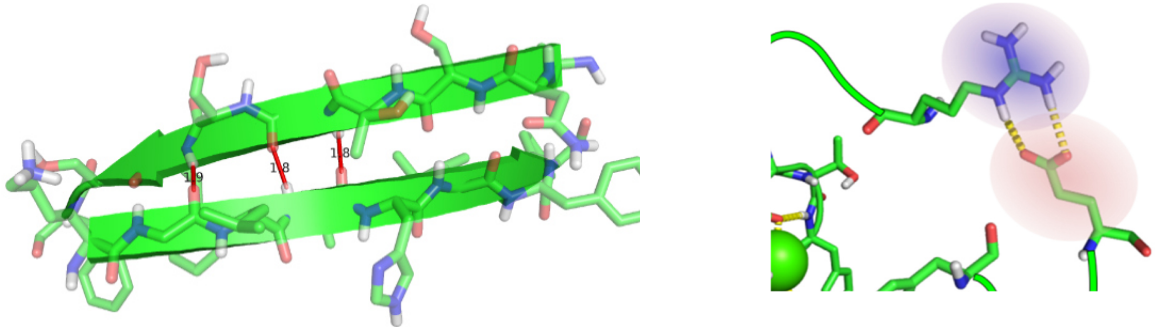
The electrostatic surface of a protein can be estimated using different techniques as the solution of the Poisson-Boltzmann equation we can estimate. These techniques are based on computing the electrostatic potential in the surface of the protein. The study of the electrostatic potential of a protein can provide information predicting putative binding sites or regions of a protein that might be embedded in the cell membrane.



Electrostatic potential of the antibody with the antigen moved from the binding site. For better visualization, the antigen was removed from the scFv and rotated such that the binding face is visible. Folding the figure in the middle would restore the original orientation of the peptide. Image from C. Zahnd et al.[73] under the terms of the (CC BY 4.0)..

HYDROGEN BONDING

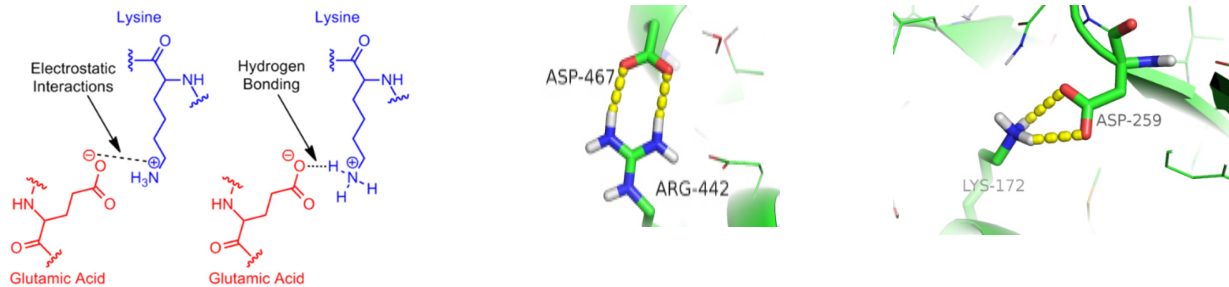
Hydrogen bonds are highly selective interactions between a hydrogen atom attached to an electronegative atom, the hydrogen bond donor, and the hydrogen bond acceptor which are also electronegative nitrogen or oxygen atoms. These interactions are one of the most important taking place in proteins because they hold the secondary structure of proteins, i.e. α -helices and β -sheets, through interactions of the oxygen and nitrogen atoms of the amide bonds of the main chains of the protein. Secondary structure motives hold the different interaction of the side chain of the amino acids are also interacting through hydrogen bonds. Furthermore, different protein subunits can form quaternary structure also interacting with these interactions. As commented previously, a rule of thumb range for the energies associated with hydrogen bonds is $6\text{-}30\text{ kJ/mol} \approx 1.4\text{-}7\text{ kcal/mol} \approx 2\text{-}12\text{ kT/e}$.



Different hydrogen bonds observed between protein atoms, left intra main chain hydrogen bonds, centre and right interaction between atoms of the side chain

SALT BRIDGE

A salt bridge is a very strong interaction due to the hydrogen bonding and ionic charge-charge bonding. Furthermore, when salt bridge occurs between certain charged amino acids as Arg-Glu, Arg-Asp or Lys-Glu, the molecular geometry allows the formation of two hydrogen bonds adding energy to the interaction.

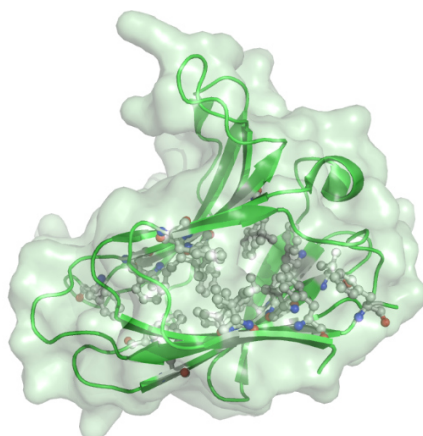


Different hydrogen bonds observed between protein atoms, left intra main chain hydrogen bonds, centre and right interaction between atoms of the side chain

HYDROPHOBIC INTERACTIONS

As previously described, proteins are polymers of amino acids. The free energy required to fold the amino acid chain into a three-dimensional structure, it has been generally accepted, that a big proportion of such energy comes from the hydrophobic effect, that is the aggregation of the hydrophobic amino acids in the core of the protein.

Then, hydrophobicity can be understood as the low affinity of nonpolar groups to interact with water molecules, mainly formed by clusters of carbon atoms. The energy cost of breaking the hydrogen bond network formed by the water solution and placing the molecule instead, could explain hydrophobicity. We can postulate that an entropic origin could play a role since temperature affects hydrophobicity. Then, hydrophobic interactions are the spontaneous tendency of nonpolar groups to adhere in water to minimize their contact with water molecules. The free energy of hydrophobic interactions is proportional to the surface area in contact [74].



Structure of a galectin protein (PDB-ID 5t7i), showing hydrophobic amino acids in the core of the protein in white.

STACKING INTERACTIONS

Aromatic-aromatic interactions are common in most molecular structures in biology. The vast majority of biomolecules contain aromatic substituents and recognition by proteins is often dominated by aromatic-aromatic interactions also called π - π interactions. Parallel and T-shaped π - π interactions are the most commonly found orientation in the interaction between aromatic groups of amino acids. Furthermore, cation- π interactions are interaction between the face of an aromatic π system and an adjacent cation can be found [75]. These interactions are attractive non-bonded interactions between aromatic rings where small charges favor parallel-displaced geometries and large partial charges favor T-shaped structures.

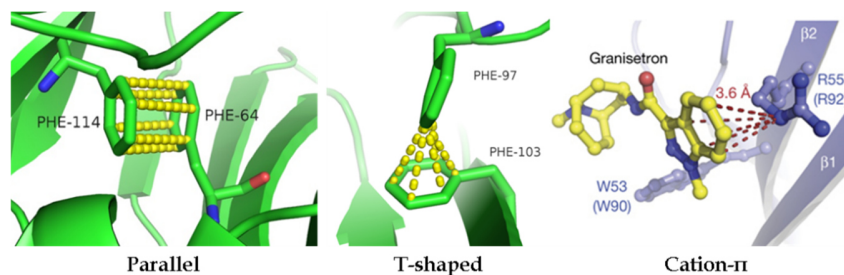


Figure showing examples of stacking interactions. Image of cation- π from D. Kesters et al. [76] (CC BY 4.0)..

PROTEIN HYDRATION, WATER AND IONS

In the experimental structure of proteins, is common to find some water or metallic molecules mediating interaction between amino acids, or even between amino acids and other protein or ligands. Furthermore, surrounding the protein, there is hydration shell around that is important for the activity of the protein, in fact most proteins lack biological activity in the absence of sufficient hydrating water. In proteins structure, we can usually find isolated solvent molecules mediating or coordinating the interaction through hydrogen bonds or pure charge-charge interactions.

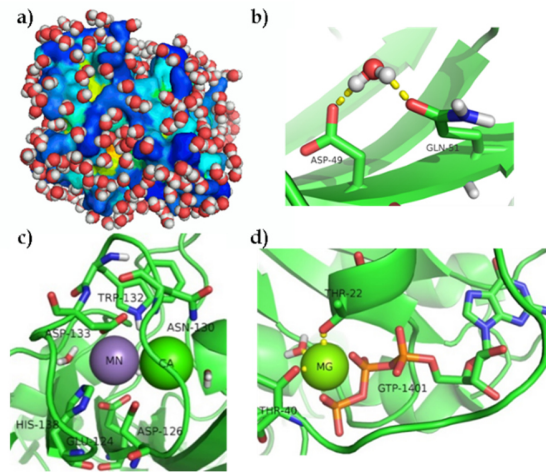
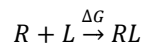


Figure showing, a) example of a protein surrounded by the hydration water shell, b) double hydrogen bond between a water molecule and two amino acids (PDB-ID 4t7i), c) metallic interaction with manganese and calcium ions (PDB-ID 4wv8), and d) crystal structure of human small GTPase Rab7 bound to GTP showing the coordination with a Magnesium ion.

BINDING BETWEEN MOLECULES

BINDING FREE ENERGY

The thermodynamics under the binding free energy (ΔG) process during the formation of a molecular complex between two partners can be described as the reaction between two molecules, one considered as the receptor "R" and ligand "L" [77],



ΔG is calculated from the ratio of the bound and unbound state probabilities,

$$\Delta G = -k_b T \ln \frac{P(R+L)}{P(RL)} = -k_b T \ln \frac{P_{Unbound}}{P_{Bound}}$$

where $k_b T$, P_{Bound} and $P_{Unbound}$ are the Boltzmann constant, temperature, and the probabilities of the bound and unbound state, respectively. Each probability is calculated by integrating an energy function at each state.

However we can use other expressions as the change in the Gibbs free energy of the system that occurs during a reaction, defined as the change in the enthalpy and entropy of the system at a certain temperature T ,

$$\Delta G = \Delta H - \Delta S$$

where enthalpy ΔH can be approximated as the difference in the potential energy between the bound and the unbound state,

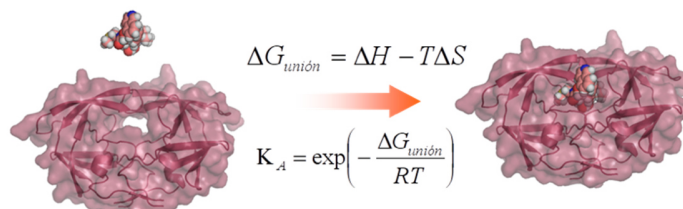
$$\Delta H = U(RL) - (U(R) + U(L))$$

Moreover, ΔS is the difference in entropy between the bound and unbound states. The entropy of a system (S) is a measure of the number of configurations Ω that a thermodynamic system can have. If we assume that all of

microscopic configurations are equally probable, we can estimate it as the product between the natural logarithm of Ω and the Boltzmann constant k_b that gives energy dimensions to the entropy,

$$S = k_b \ln \Omega$$

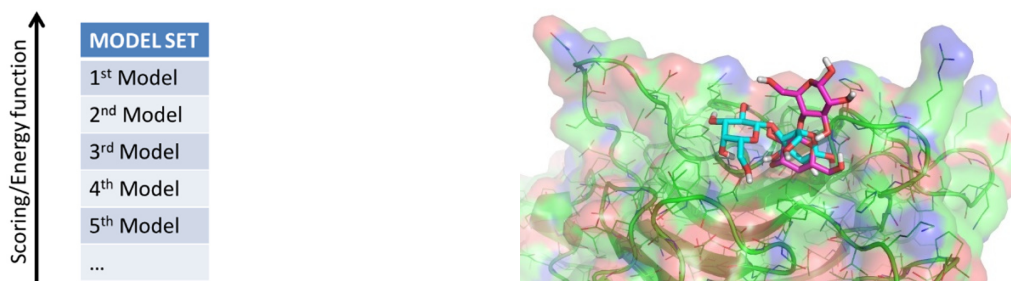
Nevertheless, the calculation of this entropic contribution represents a challenge, since the possible number of states for a complex macromolecule, as a protein or any other larger molecular system, is still a fuzzy concept with our current approaches.



Representation of a protein ligand binding together with two Gibbs Free Energy and the estimation of the binding constant.

ESTIMATION OF BINDING AFFINITY

Estimating the free energy of binding between a receptor and a ligand can be used to evaluate or score molecular models. Normally, the evaluation is carried out by using mathematical equations that provide a value that allows us to measure the strength that a ligand uses to bind a target protein, or to provide a ranked list from a certain set of binding models. That is why they are usually referred to as scoring functions or energy functions since they allow to generate such a ranking of models. It is important to consider the limitations that these approaches may have, and that, in many cases, they fail to reproduce what really occurs in nature. For example, sometimes a problem arises from evaluating vdW interactions, which may favor the contact between large surfaces when using force field based scoring functions.



Left, ranked list of models by the increasing/decreasing score from a scoring or energy function. Right, protein surface of the 4wv8 protein in complex with lactose showing two binding modes, in magenta the crystallographic ligand and in blue a wrong model.

Force Field-based Scoring Functions, decompose the ligand-receptor binding energy in the sum of individual terms as vdW, electrostatics, hydrogen bonds, hydrophobic interactions, etc. They rely on molecular mechanics parameters. For example, a simple but helpful approach may be implementing vdW and pure coulombic interactions as,

$$\Delta G_{Bind} = \sum_{ij} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} + \frac{q_i q_j}{\epsilon r_{ij}} \right]$$

Empirical Scoring Function, are based on counting the physicochemical properties of special relevance in ligand receptor binding, measuring properties as the number of ligand receptor contacts, or in the change in solvent accessible surface area. These coefficients of the scoring functions are obtained by regression functions fitted with experimental data. It is the natural extension of Hammett and Hansch ideas, Free Energy Linear Relationship, QSAR, etc. [78]

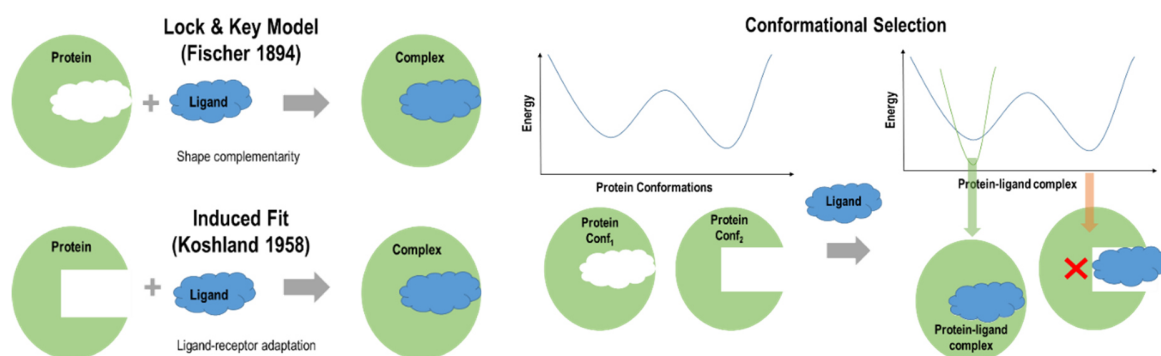
$$\Delta G_{Bind} = \Delta G_0 + \overbrace{\Delta G_{hb} \sum_{h-bonds} f(\Delta R, \Delta \alpha)}^{\text{hydrogen bonds}} + \overbrace{\Delta G_{ionic} \sum_{ionic-int} f(\Delta R, \Delta \alpha)}^{\text{salt bridges}} + \overbrace{\Delta G_{lipophilic} |A_{lipophilic}|}^{\text{lipophilic Interactions}} + \overbrace{\Delta G_{rot} NRot}^{\text{ligand entropy}}$$

Knowledge-based Scoring Functions (statistical potentials/Mean Force Potential), uses the information contained in structural databases such as the Cambridge Structural Database or Protein Data Bank, computing the frequencies of certain protein ligand interactions. These frequencies are used to derive "potentials of mean force" based in the assumption that close intermolecular interactions between certain types of atoms or functional groups that occur more frequently than one would expect by a random distribution, are likely to be energetically favorable and therefore contribute favorably to binding affinity.

$$\Delta G_{bind} = \sum_{\substack{k,l \\ \text{all pairs of ligand receptor atoms, } k \text{ and } l \text{ respectively}}} \overbrace{A_{ij}(r)}^{\text{all interactions between } i \text{ and } j \text{ atom types}} ; A_{ij}(r) = -kT \ln \frac{\overbrace{g_{ij}(r)}^{\text{probability that } i \text{ and } j \text{ will be in contact at a distance } r}}{\underbrace{g(r)}_{\text{probability of the reference state (no interaction)}}}$$

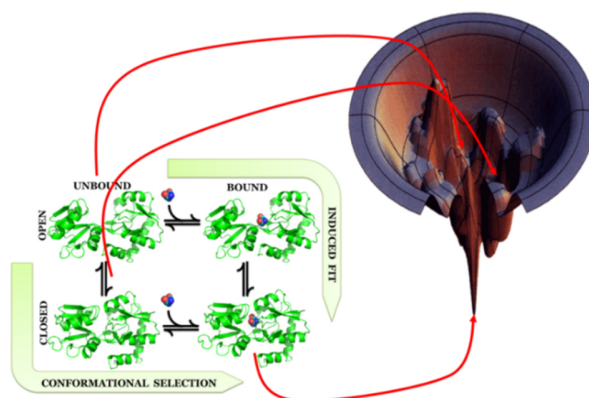
MOLECULAR ASSOCIATION MODELS

Nowadays, the understanding of the binding mechanisms between proteins has evolved [79] from the early "lock-and-key" to dynamic view of the binding event [80]. According to the induced fit model, the interaction between a protein and a ligand induces a conformational change in the protein leading to the apo/holo or bound/unbound conformations of the binding partners. Apart from the classic binding "induced fit" schema, it has been observed that, there are molecular systems that follow the "conformational selection" paradigm, where the binding partners may dynamically fluctuate between conformations, where one of the conformations is selected as the one matching with binding, i.e., the bound conformation. Then, the binding of the ligand shifts the conformational ensemble towards this bound state. Selective binding to a single conformation in the ensemble was suggested by Straub *et al.* [81].



From left to right, representation of the Lock and key, Induced fit and Conformational Selections models of molecular association.

When studying whether induced fit vs. conformational selection models should happen during the binding between two partners, we can think it as two possible scenarios. If the ligand presence produces a force that induces a conformational change of the protein, overcoming the energy barrier found between the holo and apo conformations, we may observe conformational change of the protein explained by the induced fit model. Otherwise, if the protein may fluctuate between conformations spontaneously, and the ligand binding just stabilizes one of them, we may be observing conformational change of the protein explained by the conformational selection model.

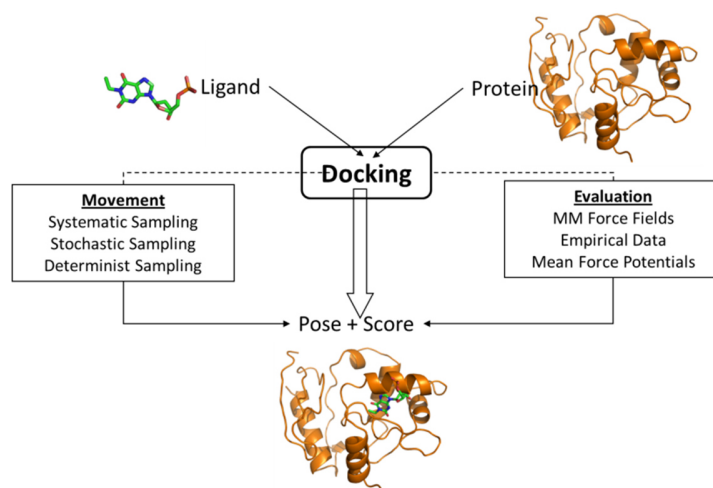


Representation of the different binding paths for induced fit and conformational selection scenarios, along with a hypothetical energy landscape marking the minimum of energy related with the conformations of the protein ligand complexes.

MOLECULAR DOCKING

Given the 3D structure of a target protein of interest and a ligand, where the ligand can be the structure of a small molecule or another protein, molecular docking is defined as the computational approach that aims to predict the experimental 3D structure of a molecular complex. A reliable docking method should account with the degrees of freedom present in a protein ligand binding event, i.e. the conformational changes, being this the one of the mayor challenges of current molecular docking approaches.

Then, docking aims predicting interactions between proteins and another molecule, normally noted as ligand. It is based on, in one hand, having an appropriate algorithm being able to sample the possible movements of the molecule, i.e. generation of “poses” within the binding site, and, in the other hand, having a fast and reliable scoring function able to perform the evaluation of the generated poses.



General schema of protein ligand main theoretical considerations.

Two types of motions describe the possible movements that the ligand can make within the binding site, rigid body motions and internal flexibility. Rigid body motions are motions where the whole ligand can move over the 3D space as a rigid unity, in the empty space left by the protein amino acids, i.e., the ligand binding site or protein pocket. Then, the internal flexibility of the ligand or the receptor that is reflected in the degrees of freedom of the binding partners, normally it is represented by rotations in the dihedral angles of rotatable covalent bonds, i.e. single bonds. [82]

Systematic search based on all possible movements, i.e. exhaustive variation on each degree of freedom, leads to a combinatorial explosion with molecules as described in previous sections:

$$N_{conf} = \prod_i \prod_j \frac{360}{\theta_{ij}}$$

Then searching for an optimum binding mode can follow different strategies. One could be moving the ligand using a stochastic approach where random changes on the degrees of freedom of the small molecule are made. Other method can use a deterministic approach, where the initial state determines the next state.

Once different binding or “poses” modes are generated, the evaluation of each generated “pose” is performed by the scoring functions previously described, i.e., quantitative characterization of the interaction based in energy or scores must be performed in order to generate a ranking of the poses.

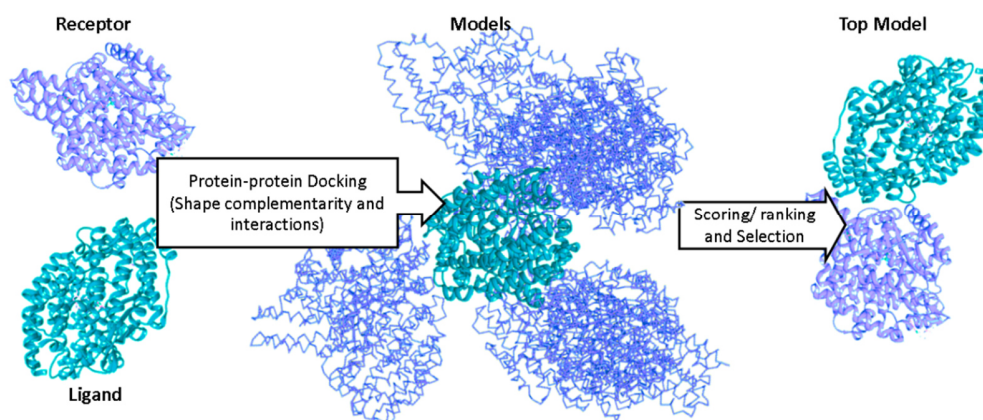
PROTEIN-PROTEIN DOCKING

The protein-protein docking may lead to misleading results since many times the interface conformations of receptor and ligand may change during the binding from the unbound to the bound state. These changes can be limited to specific residues, but sometimes involve large movements of domains, stabilization of disordered regions or domain reorganization.

Protein-protein docking approaches differ in their strategies and methodologies. For example, we can reduce the amount of possibilities by generating models focused on the specific binding domains and afterword performing superposing of the rest of the protein with this docked domains [83], [84].

Then, most protein-protein docking models apply a combination of energetic evaluation applying simple scoring functions and shape complementarity of initial models from rigid body transformations of the ligand over the receptor, and later applying optimization algorithms as energy minimization for adapting the side-chain conformations.

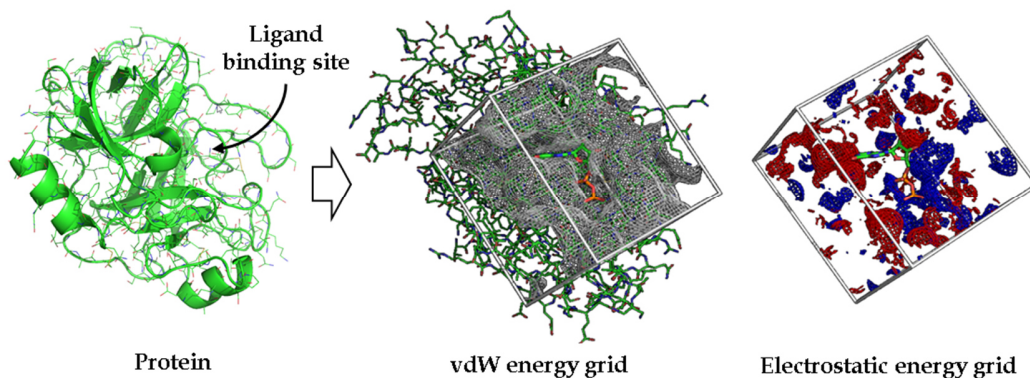
There exists several client based tools and web servers for performing protein-protein docking from which we can point the web servers with remarkable performance ClusPro [85] and PyDockWeb [86]. As a reference, ClusPro [85] combines the generation of docking models based also in shape complementarity and later side-chains refinement as described previously, later it applies a classification analysis providing a set of diverse clusters where most populated cluster may indicate a higher probability of reproducing the natural binding mode between the proteins.



Pictorial representation of the protein-protein docking process.

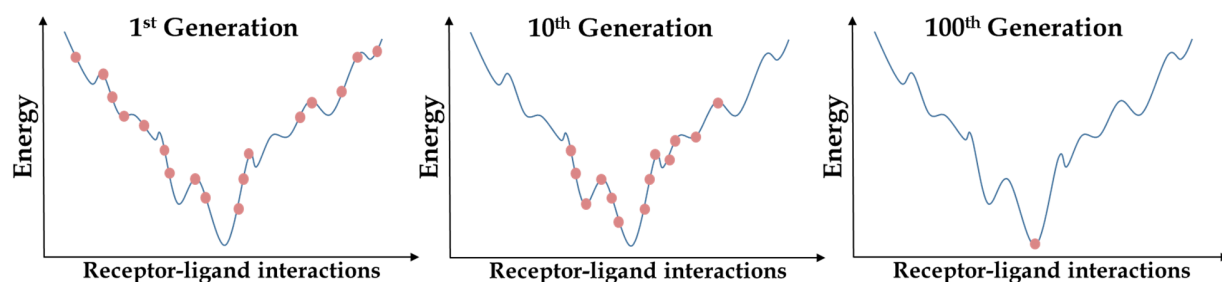
PROTEIN-LIGAND DOCKING

Since computing a full scoring or energy function can be computationally exhaustive, a common approach is to generate a set of energy grids that will be used for all poses generated by the movement algorithm [87]. These energy grids are usually defined within an octahedral box centered in the binding site of the protein, and they help to perform fast evaluation of each generated poses. Energy grids resume the space of the binding site on a set of individual points where different atom types contained in the ligand (oxygen, carbon, nitrogen, etc.) are placed and energetically evaluated, recording an energy value for a certain atom type in a certain space point. Then, when a new ligand binding mode or pose is generated by the movement algorithm, it allows for a fast evaluation of the pose by looking at the energy value from the closest grid point for the specific atom type contained in the ligand. Then the final binding energy or score results from the summation of all the energy values close to the different atoms in the ligand in the specific generated ligand pose.



Representation of a vdW grid (right) accounting for the shape/surface of the binding site, and a coulombic energy grid (right) accounting for electrostatic interactions

Having both, the capability of evaluating a certain binding pose, and the correct tool for evaluating a protein-ligand complex, different methods have been implemented to perform the docking search using deterministic approaches as minimization of random poses, or the implementation of genetic algorithms as the Lamarckian algorithm [88] implemented in the open source protein-ligand docking program Autodock [89].



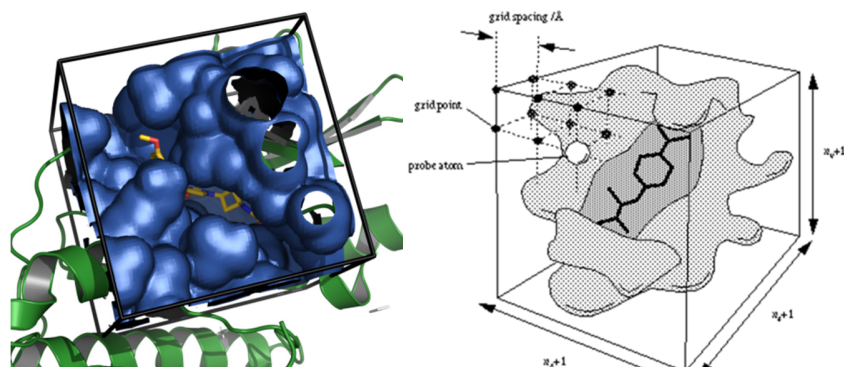
Simple representation of a genetic algorithm relating the different binding modes with its energy over the different steps of the algorithm.

DOCKING APPROACHES

There are different docking methods implemented over the most common protein-ligand docking programs or suites as the previously mentioned open source program Autodock [89], among other commercially available as glide [90].

Initially, the ligand binding site must be defined into a certain volume coinciding with the energy grids previously defined, normally a octahedral box centered in the ligand binding site of the protein when known, in order to reduce the amount of possible movements of the ligand. Sometimes we will not know where the ligand binding site of the protein is located. In such cases there are other programs and servers for the prediction of ligand binding

sites as for example fpocket [91], which looks for protein cavities that could potentially represent ligand binding sites, due to their physicochemical environment and topology.

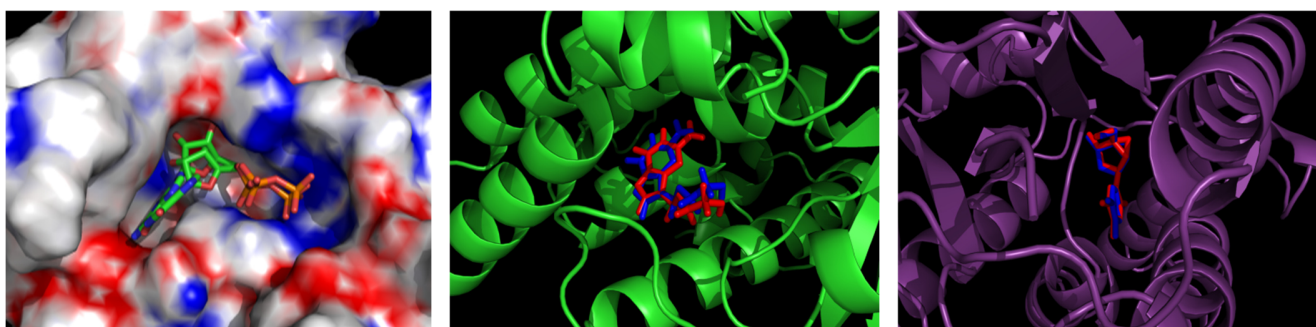


Left, image of an octahedral box defining the binding site for computing the energy grids and volume to perfume the ligand movements, where the protein is represented in green, the surface at the binding site in dark blue, and the buried docking pose of the ligand in yellow. Image from Autodock Vina plugin web [92]. Right, representation of the octahedral grid box, representing the equally spaced grid points, an example of a probe atom to compute energies for a certain atom type, the protein surface and the buried ligand. Image from Autodock tutorial web [89].

The most simplistic scenario emulated by many docking approaches will consider the protein as a rigid entity. This is valid when experimental information is available for the ligand bound conformation of the protein assuming that new ligands will interact in that conformation, as for example when the experimental structure of the protein is known in bound conformation. Nevertheless, normally we are able to consider ligands as a flexible entities defining its degrees of freedom based on the dihedral angles of rotatable bonds.

Other approaches requiring higher computation capabilities, can offer the description of different levels of protein flexibility, we can simulate small movements of side chains of the protein, or simulate larger conformation changes that may happen during the binding.

Experimental structures of proteins often present well-defined ligand pockets or cavities. In such cases, docking algorithms normally succeed in predicting the bioactive ligand pose.



Three examples of docking succesfull cases where there proposed model matches the experimental structure.

DRUG DISCOVERY AND RATIONAL DRUG DESIGN

Drug discovery make use of different experimental and computational approaches depending on the complexity of the problem and resources available. Methods range from the use of pure High Throughput Screening campaigns where individual assays are performed with thousands of compounds in parallelized experiments demanding many experimental resources, or combining computational approaches as the so-called Virtual Screening that can be computed at ligand level or based on the protein receptor, reducing costs at different steps of the search.

3D structure	Ligand	
	unknown	known
Receptor		
Unknown	High Throughput Screening (HTS)	Pharmacophore-based Virtual Screening QSAR
Known	Receptor-based Virtual Screening <i>de novo</i> methods	Docking Virtual Screening

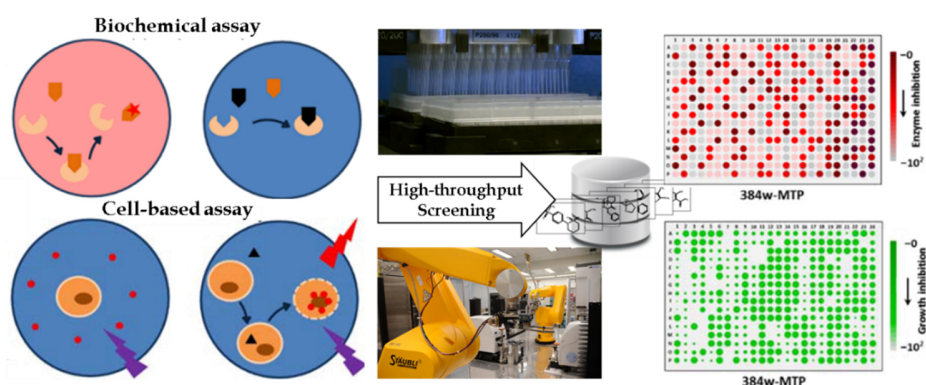
Table showing the possible scenarios for applying molecular Drug Discovery and Rational Drug Design for predicting small molecule ligand binders to a target receptor.

HIGH THROUGHPUT SCREENING

High-Throughput Screening (HTS) technology combines robotics, miniaturization, sophisticated biological assays and chemistry, and sophisticated software and database implementations. High-throughput assays are performed in 96 or 384 well plates, where a chemical library of chemical compounds is tested either in cellular or proteins activity assays. The parallelization of the assay is the main remarkable feature from this technique allowing the analysis in the range of thousand compounds requiring robotics automation that supposes an important economical investment.

A standard HTS procedure, consists on first the initial hit identification over a chemical library, then the confirmation of the top potential hits by at least triplicate assay, and finally the estimation of the molecular activity as IC₅₀ or EC₅₀ values with at least 10 point dose-response assay to check the potency of the compounds.

HTS assay examples comprehend from receptor-binding biochemical experiments targeting protease, kinase, phosphatase, lipase, others, bacterial growth, cell-based reporter gene, cell growth, cell viability, Cytochrome P450 inhibition, protein production by cells ELISA [93]–[96].



Schema showing the steps in High Throughput Screening as biochemical or cellular assay set up, robotics automation of the assay for testing large chemical libraries, and result interpretation. Adapted from C. Regnault [93] and Wikimedia Commons [7] (CC BY 4.0)..

VIRTUAL SCREENING

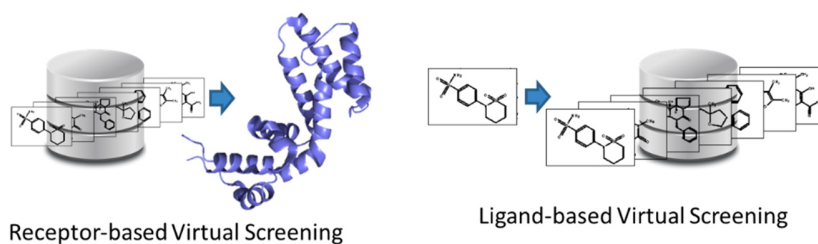
Usually the objective of applying computational approaches in the drug discovery process is to create a protocol that reduce the number of assays at different steps of the experimental protocol finally allowing us to distinguish between active and inactive molecules.

Virtual Screening is of great help in experimental laboratories and High Throughput Screening campaigns since it allows focusing in a reduced set of compounds from large chemical libraries with an incremented probability of finding active compounds in the experimental assay. Then Virtual Screening can lead to a reduction of time and costs in the search process. It is also important to notice that Virtual Screening relies in theoretical approaches and can lead to an important rate of false positive ratio.

Virtual Screening approaches are divided in Receptor-based, where prior knowledge of the target protein is known, and Ligand-based where prior knowledge of a chemical compounds able to modulate a certain activity is known. In the best of the scenarios where a previous protein-ligand interaction with an already known 3D structure is known, both ligand and structure based can be also combined.

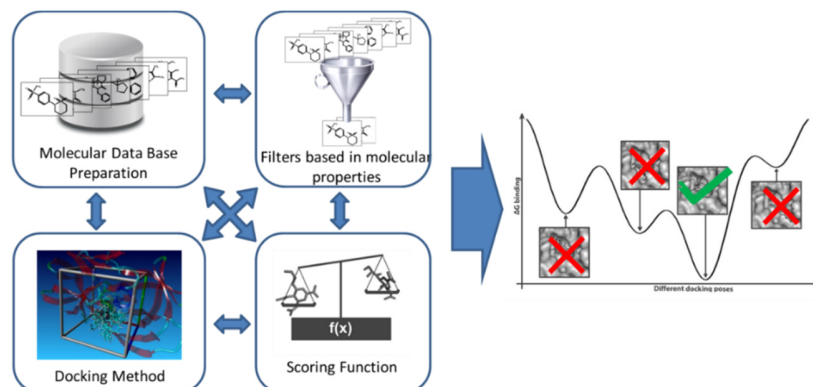
Receptor based is performed by computing the individual docking over a database of compounds. Since the computational cost associated can be high, it is normally performed in parallel with the help of computer clusters over hundreds even thousands of individual computing nodes.

Ligand based approach can differ, but as example we could apply fingerprint comparisons to filter compound with a certain chemical substructure, or pharmacophore-based comparisons to seek for compounds with certain physicochemical properties.



Representation of the main approaches used in Virtual Screening.

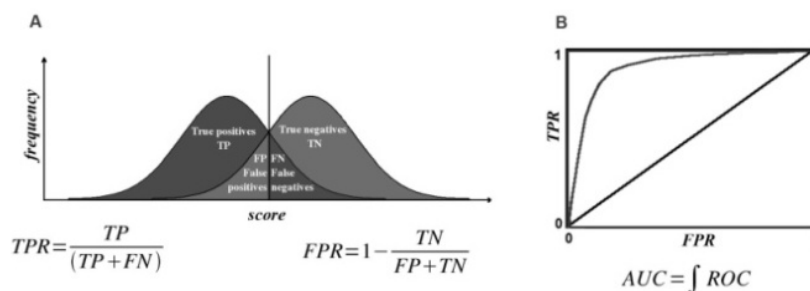
The general strategy followed in a Receptor-based Virtual Screening experiment can consist in some or all of the following experiments. First, the study and preparation of the target protein and the preparation of the chemical library into a molecular database containing prepared and optimized ligand structures. Second, prescreening or filtering the database by molecular weight, hydrophobicity, or applying Ligand-based Virtual Screening in case some information about other active compounds is already known. Third, screening of compounds where thousands or millions of individual dockings are performed. Fourth, re-scoring or ranking docking results leaving active compounds on the top of the list. Fifth, post screening filters and visual inspection. And sixth, the finally experimental assay to of the top compounds, the number of compounds tested will depend on the availability.



Representation of the components and combinations in Virtual Screening protocols.

EVALUATING SCREENING PROTOCOLS

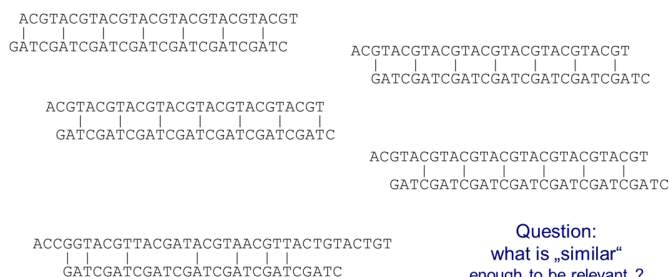
Accurate analysis of the results of a HTS campaign is crucial to drug discovery. Misleading as false positive or false negative values may rise at different steps of the assay protocol, and the analysis should identify those false positive that may bias the results. The evaluation of virtual screening protocols is usually performed by the study of the false positive, true positive rates (FPR and TPR) and the analysis of Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC) curves. ROC curves are created by plotting the TPR against FPR at various percentages of the database tested. Then ideally, the AUC value should be one when all positives are ranked at the top of the list by the Virtual Screening experiment.



A, hypothetical distribution of True Positive and False Positive Rates (TPR and FPR). B, representation of a AUC ROC curve.

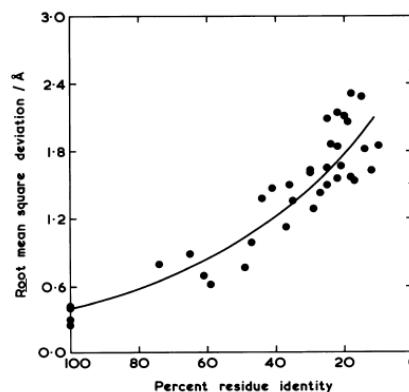
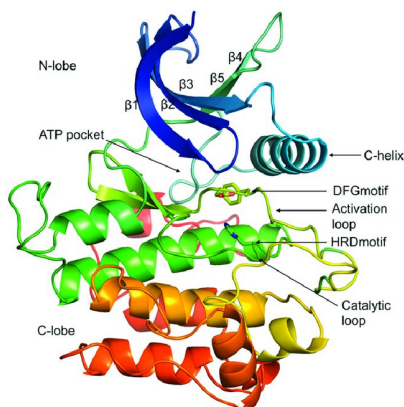
PROTEIN SIMILARITY

Aligning protein and DNA sequences has been one of the tools that have helped to understand evolution by measuring the similarity between genomes or proteomes of different organism giving a measure to the homology between proteins. Aligning or matching a gene or a protein sequence is achieved by dynamic programming, which should provide an optimal alignment between two sequences allowing conserved mutations to be paired, or the insertion or deletion of sequence segments by inserting gaps. Needleman- Wunsch Algorithm [97] for global alignment or Smith & Waterman Algorithm [98] for local alignment are two examples.



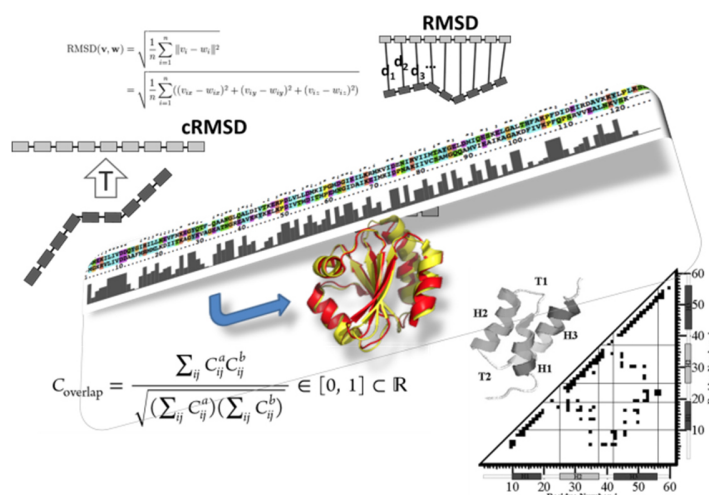
DNA sequence alignment.

Comparing the sequence of related proteins can identify how changes at the sequence level may affect to the function of the protein. In addition, protein function is completely related to the three-dimensional structure of proteins. It has been shown that many homologous proteins share their structural features, as for example protein kinases, that have a characteristic catalytic domain, which transfers the phosphate from the nucleoside to an amino acid of the phosphorylated protein. This catalytic domain is highly conserve among the kinase families, which means that they all have a similar nucleotide binding site with the amino acids required for the catalysis in their specific places. Furthermore, it has been also shown that the similarity between homologous proteins can be higher in terms structure than the similarity in terms of the sequence.



Left Structure of a typical protein kinase domain displaying ATP binding site and conserved elements around it (INSR kinase, PDB ID code 1GAG). Image from V. Modi et al. [99] under the terms of the (CC BY 4.0). Right, The relation of residue identity and the RMSD. deviation of the backbone atoms of the common cores of 32 pairs of homologous proteins, i.e. the relation between sequence similarity and structure similarity. Image from Chothia & Lesk [100] (CC0 1.0)

Then it is of mayor interest having the right tools to assess whether two proteins are related in terms of structure. For this purpose, there are modelling approaches that provide an appropriate estimation of the similarity between two proteins considering both, the three dimensional arrangement of the protein as well as the alignment of the amino acid sequence. From those methods, we will review two simple approaches, one based on the use of RMSD, and other based on the comparison of the internal interactions or contacts between the amino acid of the protein.



Summary/Representation of some protein similarity techniques.

SEQUENCE ALIGNMENT

Nowadays, there are several databases that store the information of all know proteins sequences, and it is of great help to have tools to compare different gene and protein sequences. This will help us to identify evolutive relationships between homologous genes. The most common way to compare two sequences is to perform a sequence alignment. There is not a unique sequence alignment but an optimal alignment, this means that we can perform several alignments between two sequences, but one will better than the other depending on certain parameters that will help us to evaluate it.

In one hand, we can measure the sequence identity by summing up a score of one for each identic nucleotide. Or measure the sequence similarity by giving different scores for the different nucleotides aligned. For example, when aligning DNA sequences, we can assume that aligning cytosine with cytosine or guanine with guanine is highly favorable, whereas aligning adenine with thymine is favorable but not that much, penalizing when other nucleotides are aligned, or we even have to introduce gaps in the sequence.

TGAAGTA-CT
TCATGTACACT

Identity: 1+0+1+0+1+1+1+0+1+1+1=8

Similarity: 1-2+1-1+2+1+1-4+1+2+1=3

Similarity Score:	
C → C, G → G	+2
A → T, T → A	+1
C → G, G → C	-1
Other	-2
Gap	-4

Representation of two different ways to evaluate a sequence alignment, first by sequence identity and second by sequence similarity. Figure adapted from [101].

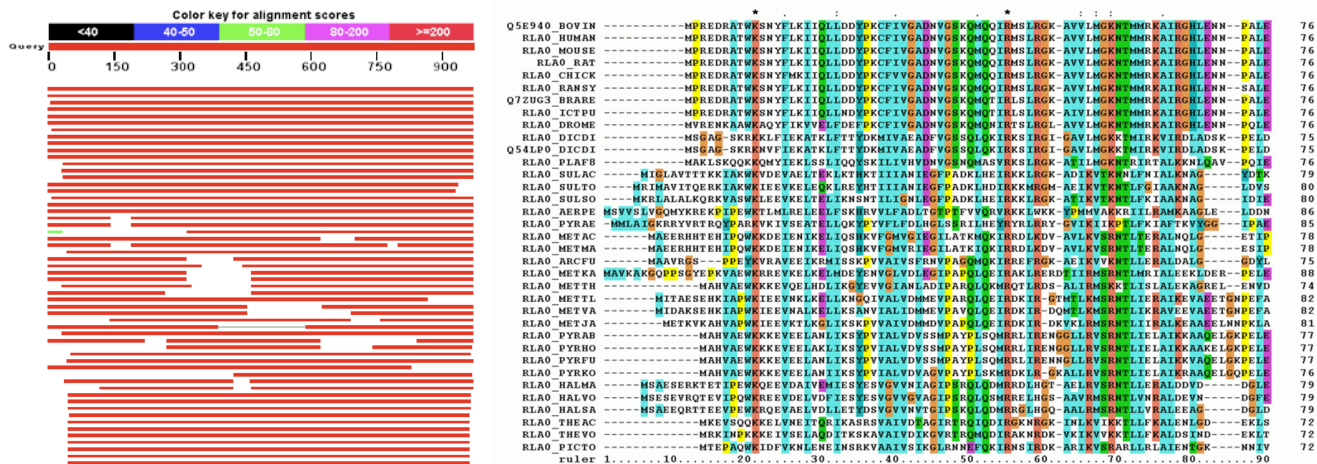
After defining what a sequence alignment is for a nucleotide sequence, we can extend the same approach for protein sequences, searching for the optimal alignment by using different algorithms previously referenced, but defining new similarity scores for the different amino acids substitutions in the protein sequence. This leads to the definition of substitution matrices as the Point Accepted Mutation (PAM) matrix or the BLOck SUBstitution Matrix (BLOSUM) [102].

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	5	-4	-2	-1	-4	-2	-1	0	-4	-2	-4	-4	-3	-6	0	1	1	-9	-5	-1	-1	-2
R	-4	8	-3	-6	-5	0	-5	-6	0	-3	-6	2	-2	-7	-2	-1	-4	0	-7	-5	-4	-2
N	-2	-3	6	3	-7	-1	0	-1	-1	-3	-5	0	-5	-6	-3	1	0	-6	-3	-5	5	-1
D	-1	-6	3	6	-9	0	3	-1	-1	-5	-8	-2	-7	-10	-4	-1	-2	-10	-7	-5	2	-3
C	-4	-5	-7	-9	9	-9	-9	-6	-5	-4	-10	-9	-9	-8	-5	-1	-5	-11	-2	-4	-8	-9
Q	-2	0	-1	0	-9	7	2	-4	2	-5	-3	-1	-2	-9	-1	-3	-3	-8	-8	-4	-1	5
E	-1	-5	0	3	-9	2	6	-2	-2	-4	-6	-2	-4	-9	-3	-2	-3	-11	-6	-4	2	5
G	0	-6	-1	-1	-6	-4	-2	6	-6	-6	-7	-5	-6	-7	-3	0	-3	-10	-9	-3	-1	-3
H	-4	0	1	-1	-5	2	-2	6	8	-6	-4	-3	-6	-4	-2	-3	-4	-5	-1	-4	0	1
I	-2	-3	-5	-4	-5	-4	-6	-6	7	1	-4	1	0	-5	-4	-1	-9	-4	3	-4	-4	-3
L	-4	-6	-5	-8	-10	-3	-6	-7	-4	1	6	-5	2	-1	-5	-6	-4	-4	4	0	-6	-4
K	-4	2	0	-2	-9	-1	-2	-5	-3	-4	-5	6	0	-9	-4	-2	-1	-7	-7	-6	-1	-2
M	-3	-2	-5	-7	-9	-2	-4	-6	-6	1	2	0	10	-2	-5	-3	-2	-8	-7	0	-6	-3
F	-6	-7	-6	-10	-8	-9	-9	-7	-4	0	-1	-9	-2	8	-7	-4	-6	-2	4	-5	-7	-9
P	0	-2	-3	-4	-5	-1	-3	-3	-2	-5	-5	-4	-5	-7	7	0	-2	-9	-9	-3	-4	-2
S	-1	-1	-1	-1	-3	-2	0	-3	-4	-6	-2	-3	-4	0	5	2	-3	-5	-3	0	-2	-1
T	-1	-4	0	-2	-5	-3	-3	-3	-4	-1	-4	-1	-2	-6	-2	2	6	-8	-4	-1	-1	-3
W	-9	0	-6	-10	-11	-8	-11	-10	-5	-9	-4	-7	-8	-2	-9	-3	-8	-13	-3	-10	-7	-10
Y	-5	-7	-3	-7	-2	-8	-6	-9	-1	-4	-4	-7	-7	4	-9	-5	-4	-3	9	6	-4	-7
V	-1	-5	-5	-4	-4	-4	-3	-4	3	0	6	0	5	-3	-3	-1	-10	-5	6	-4	-4	-2
B	-1	-4	5	5	-8	-1	2	-1	0	-4	-6	-1	-6	-7	4	0	-1	-7	-4	5	1	-2
Z	-1	-2	-1	-2	-9	5	5	-3	1	-4	-4	-2	-3	-9	-2	-2	-3	-10	-7	-4	1	5
X	-2	-3	-2	-3	-6	-2	-3	-3	-3	-3	-4	-3	-3	-5	-3	-1	-2	-7	-5	-2	-2	-3

Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																		
Arg	-1	5																	
Asn	-2	0	6																
Asp	-2	-2	1	6															
Cys	0	-3	-3	-3	9														
Gln	-1	1	0	0	-3	5													
Glu	-1	0	0	2	-4	2	5												
Gly	0	-2	0	-1	-3	-2	-2	6											
His	-2	0	1	-1	-3	0	0	-2	8										
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4									
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4								
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5							
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5						
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	4	7				
Ser	1	-1	1	0	0	1	0	0	-1	-2	-2	0	-1	-2	-1	4			
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5		
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	-2	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1

Example of PAM and BLOSUM substitution matrices, left and right respectively, for scoring the sequence alignment between protein sequences. Images from Wikimedia Commons [7] under the terms of the (CC BY 4.0).

Furthermore, comparing the amino acid sequences of proteins or the nucleotides sequences, also called a query with a library or database of sequences, can help identifying database sequences that resemble the query sequence above a certain threshold. This is of mayor interest helping to stablish relations between individuals. BLAST [103] (basic local alignment search tool)[2] is an algorithm and program for performing such sequence alignment using different dynamic programming algorithm to find the optimal alignment in such a scenario. This approach is noted as multiple sequence alignment.



Left, screen shot of BLAST results for protein CCDC132. Right, First 90 positions of a protein multiple sequence alignment of instances of the acidic ribosomal protein P0 (L10E) from several organisms. Images from Wikimedia Commons [7] under the terms of the (CC BY 4.0).

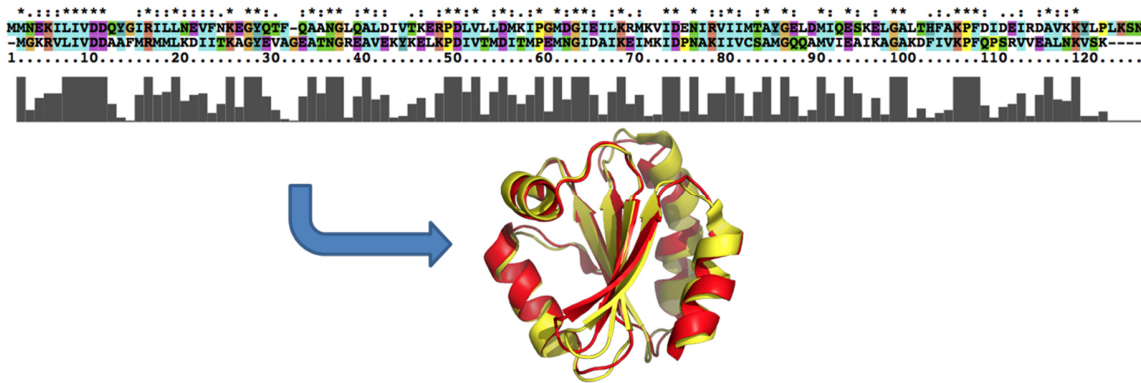
STRUCTURAL ALIGNMENT OF PROTEINS

As defined in previous section, $RMSD$ or $cRMSD$ can only be computed between molecular structures having exactly the same number of atoms. However, when two proteins have a relation at the sequence level, we can estimate the difference between the three-dimensional conformations of the proteins using its sequence alignment. This alignment will provide a relation between pair of amino acids and then we can compute the $cRMSD$ between the α -carbons of each pair of amino acids. This will be called structural alignment or structural superposition. Then, given to proteins P_A and P_B with a certain level of homology and an optimal sequence alignment we can use the $cRMSD$ we previously defined for the atoms any two conformation of a molecule, but know applied between α -carbons of the sequence alignment between the two proteins. Then, note the α -carbons of the amino acid pairs related by the sequence alignment as $\{C\alpha_1^A, C\alpha_2^A, \dots, C\alpha_n^A\}$ and $\{C\alpha_1^B, C\alpha_2^B, \dots, C\alpha_n^B\}$, and lets define the structural alignment between the two proteins as:

$$cRMSD(P_A, P_B) = \min_T \left(\frac{1}{n} \sum_{i=1}^n d_{eucli}^2(C\alpha_i^A, T(C\alpha_i^B)) \right)$$

where $T(C\alpha_i^B)$ is the rigid body transformation applied to P_B that minimizes the $RMSD$, i.e., the applying the optimal translation and rotation that superpose both P_A and P_B providing the minimum $RMSD$ value.

This approach provides simple tool for the comparing proteins with a certain degree of sequence similarity, as well as the possibility of visualizing that structural superposition for visual inspection, helping to establish structural relationships between proteins.



Representation the similarity between two proteins, $cRMSD$ between the α -carbons paired by sequence alignment.

PROTEIN CONTACT MAP

A protein contact map is a representation that allows us to identify which amino acids are at a close distance within the structure of a protein of n amino acids. Then we can represent the protein structure as binary $n \times n$ matrix C , where each position C_{ij} , will describe whether two amino acids i and j are in contact or not. The ij element of the matrix is will be 1 if the two residues are closer than a predetermined contact threshold and 0 otherwise.

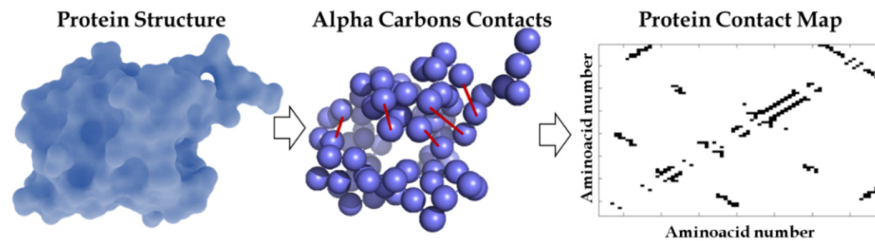
The distance between the two amino acids can be represented as the distances between α -carbons, and the contact threshold may be defined between 6-12 Å. Then,

$$C = \begin{pmatrix} C_{1,1} & \dots & C_{1,n} \\ \vdots & \ddots & \vdots \\ C_{n,1} & \dots & C_{n,n} \end{pmatrix}$$

where the contact between atoms i and j will be given by

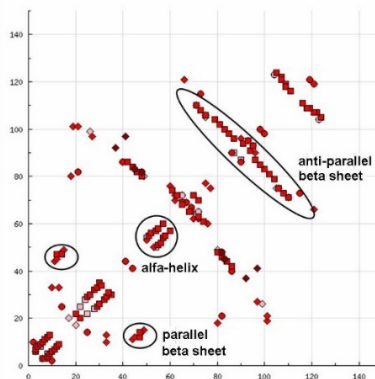
$$C_{i,j} = \begin{cases} 1 & \text{if } d(C\alpha_i, C\alpha_j) \leq \text{Contact Threshold} \\ 0 & \text{if } d(C\alpha_i, C\alpha_j) > \text{Contact Threshold} \end{cases}$$

The contact protein contact map is usually represented as a colored matrix representing the 1 values in black and 0 values in white, where subsequent amino acids, that will always be in contact, are also plotted in white for a matter of simplicity, leaving the map clearer. Note that sometimes, we only represent the upper or lower triangular matrix to avoid redundancy, since $C_{i,j} = C_{j,i}$.



Representation of the protein contact map of a protein. Note that axis represents the amino acid number and black dots represent the contact between amino acids.

A protein contact map is helpful to identify secondary structure motifs in the protein sequence. α -helixes will be in contact with residues near them, because an amino acid belonging to an α -helix makes a hydrogen bond with an amino acid 3-4 times ahead as explained in previous sections. This lead to a characteristic line parallel to the diagonal. Then, parallel β -sheets will be characterized by a line parallel line to the diagonal since an increasing numbering of amino acids in the protein sequence will interact through hydrogen bonds with and increasing numbering of the amino acids from another region of the sequence. Finally, anti-parallel β -sheets will be perpendicular to the diagonal since an increasing numbering of amino acids in the protein sequence will interact through hydrogen bonds with and decreasing numbering of the amino acids from another region of the sequence.



Representation of the protein contact map of a protein, labeling a region with antiparallel β -sheets, parallel β -sheet and an α -helix. Image from Wikimedia Commons [7] under the terms of the (CC BY 4.0).

CONTACT OVERLAP:

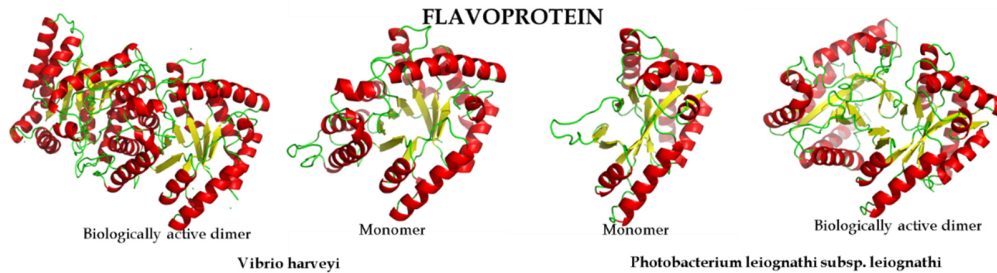
Simulation and sampling techniques can produce millions of conformations that usually need the estimation of how similar are to each other. For these purpose cRMSD can be applied but when dealing with elevated number of structures it can be too demanding in terms of computation. Then, other approaches have been developed. Given the structure of a protein in two different conformations and the contact matrixes of each of them C^a and C^b , Contact Overlap is defined by

$$C_{overlap} = \frac{\sum_{ij} C_{ij}^a C_{ij}^b}{\sqrt{(\sum_{ij} C_{ij}^a)(\sum_{ij} C_{ij}^b)}} \in [0,1]$$

As well as the cRMSD, we can compare the structure of two proteins with a certain degree of homology by defining the contact matrices of two proteins "a" and "b" considering only the amino acids that have been paired by the sequence alignment.

STRUCTURAL CLASSIFICATION OF PROTEINS

Studying the evolution of proteins brings questions as why new proteins have been emerging along time. There have been defined different mechanisms by which the structure of proteins have evolved along time as vertical transference, horizontal transference, sequence duplication in the organism or fusion of sequence fragments in the organism. Understanding the relation between sequence, structure and function of proteins will allow us to define methods to predict the function of proteins based on its sequence and structural information, giving hints to what events are predominant over the evolution of the living organisms.



Representation of the structures of flavoprotein from a bacterial luciferase protein from Vibrio harveyi and a non-fluorescent flavoprotein from Photobacterium leiognathi subsp. leiognathi. The second monomeric protein structure experience the deletion on the tertiary structure leading to a significant difference in the structure of the biologically active dimer.

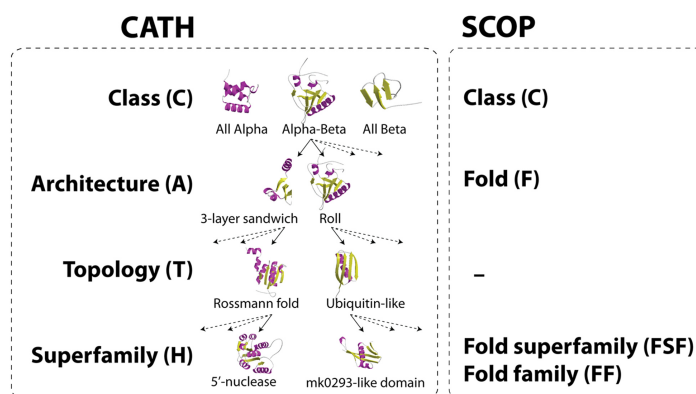
Then, a classification scheme for the structure of known proteins is of mayor interest for structural biology. There are different standard classification methods describing different classification levels for each protein. Some of these methods are performed manually or automatically. Some of them have been converted into web resources as the manual classification SCOP, or the semi manual classification of CATH or the fully automatic structural.

STRUCTURAL CLASSIFICATION DATABASES: SCOP AND CATH

The Structural Classification of Proteins (SCOP) database is a largely manual classification of protein structural domains based on similarities of their structures and amino acid sequences. A motivation for this classification is to determine the evolutionary relationship between proteins. There are several levels of SCOP classification: 1) Class: types of folds, e.g., β -sheets, 2) Fold: the different shapes of domains within a class, 3) Superfamily: the domains in a fold are grouped into superfamilies' that have at least a distant common ancestor, and 4) Family: The domains in a superfamily are grouped into families, which have a more recent common ancestor.

Proteins with the same shapes but having little sequence or functional similarity are placed in different "superfamilies". They are assumed to have only a very distant common ancestor. Proteins having the same shape and some similarity of sequence and/or function are placed in "families" and are assumed to have a closer common ancestor. Proteins having the same shape and some similarity of sequence and/or function are placed in "families" and are assumed to have a closer common ancestor.

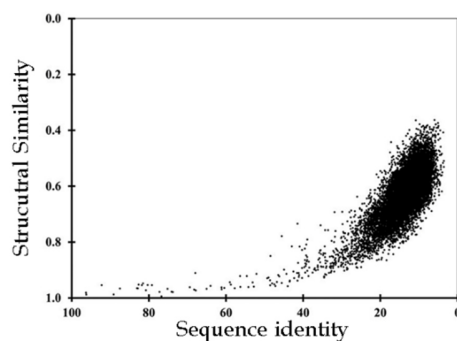
The Class, Architecture, Topology (fold family) and Homologous superfamily (CATH) database is a hierarchical domain classification of protein structures of the Protein Data Bank. Only structures solved to resolution better than 4.0Å are considered. Protein structures are classified using a combination of automated and manual procedures. There are four major levels in this hierarchy: class, architecture, topology (fold family) and homologous superfamily. The class is determined according to the secondary structure composition and packing within the structure, where three major classes are recognized as mainly- α , mainly- β and α - β . Architecture describes the overall shape of the domain structure as determined by the orientations of the secondary structures but ignores the connectivity between the secondary structures, which is assigned manually using a simple description of the secondary structure arrangement, e.g. barrel or 3-layer sandwich. Topology (Fold family), where structures are grouped into fold groups at this level depending on both the overall shape and connectivity of the secondary structures, that is done using the structure comparison algorithms. Finally, homologous superfamily groups together protein domains that are thought to share a common ancestor and can therefore be described as homologous, where similarities are identified either by high sequence identity or structure comparison.



Hierarchy of the CATH structural classification system compared to corresponding SCOP levels. Image from S. Bukhari et al. [104] (CC BY).

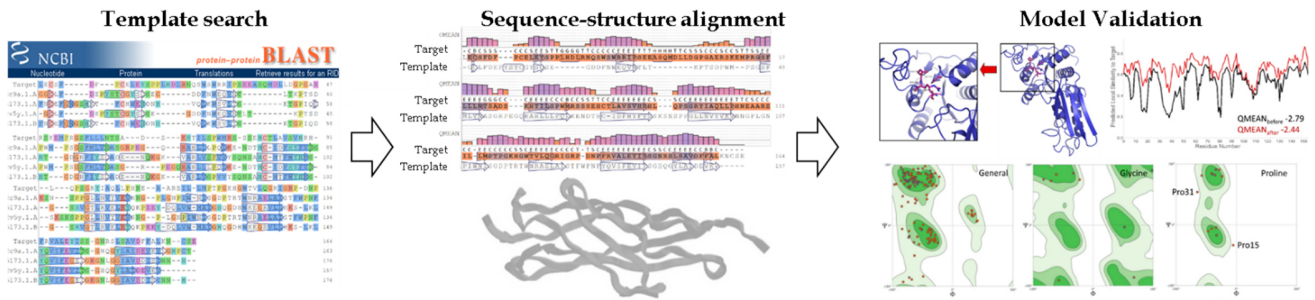
HOMOLOGY MODELLING

Homology modelling, is a molecular modelling technique based on the reconstruction of the structure of a protein at atomic detail from the amino acid sequence (the “target”) and an experimental three-dimensional structure of a protein that is homologous (the “template”). The basic idea behind Homology Modelling is that, since a protein structure is defined by its amino acid sequence, closely related sequences adopt highly similar structures, and distantly related sequences may still fold into similar structures, depending on its evolution process. Three-dimensional structure of proteins from the same family is more conserved than their primary sequences.



Correlation between sequence identity and structural similarity in the HADSF core domain. Point notes one protein pair with percent sequence identity value plotted on the x-axis and the fTM score plotted on the y-axis. Adapted from C. Pandya et al. [105] (CC BY 4.0).

The general scheme of the construction of a homology model consists in different steps described as followed. First, searching for related proteins to the target sequence with known 3D structure, by searching for homologous proteins contained the PDB using the multiple sequence alignment BLAST explained before. Second, we must select templates taking into account the sequence similarity, phylogenetic similarity in terms of solvent, pH, whether the 3D structure is in its bound or unbound conformation due to the presence of ligands or quaternary interactions. The resolution of the selected template will determine the quality of the homology model. Third, there are various computational approaches for aligning the target sequence with template structures and building a model of the target protein. Some approaches may use the information from the template structure, i.e., fitting the amino acid sequence of the query sequence into the 3D structure of the template structure, perform Molecular Mechanics calculations for optimizing the structure or applying different levels of similarity between the template and target, etc. Forth, we should perform an evaluation of the model, as checking active site residues, revising the stereochemistry and protein binding sites, etc.



Representation of the different steps performed to generate a homology model. Image adapted from Y. Haddad i et al. [106] (CC BY).

EXPLORING THE MOTIONS OF BIOMOLECULES

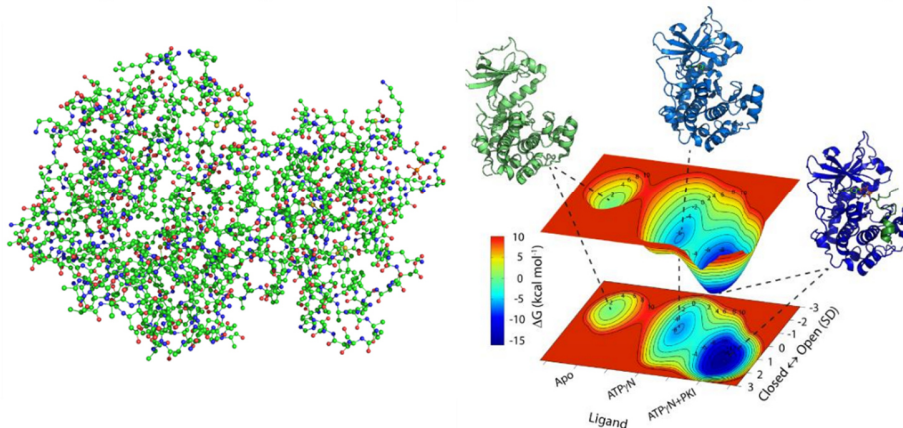
When complex biomolecules are at the interior or surroundings of cells, they are normally surrounded, either by solvent at different pH or in hydrophobic environments, depending on the organ or tissue where they belong. This, together with being at a certain temperature make molecular systems to be not static. Molecules explore complex energy landscapes than can be studied under different definitions of energy, as the ones described previously in this course by molecular mechanics force fields.

FORCE FIELDS AND MOLECULAR MECHANICS

Molecular Mechanics is an atomic detail computer simulation, where we define a force field-based potential energy function for a molecular system of atoms and use it to explore the energy surface, or free energy landscape, defined by the potential function.

Remember that, as explained in previous sections, a molecular force field is a collection of potential energy equation and parameters for the different atom types, partial charges, vdW parameters, force constants and ideal values for bond lengths, bond angles and dihedrals that define the molecular system of atoms in the molecule. Then the total energy of a molecule is divided into several terms. Potentials energy functions are calculated independently and summed to give the total energy of the molecule.

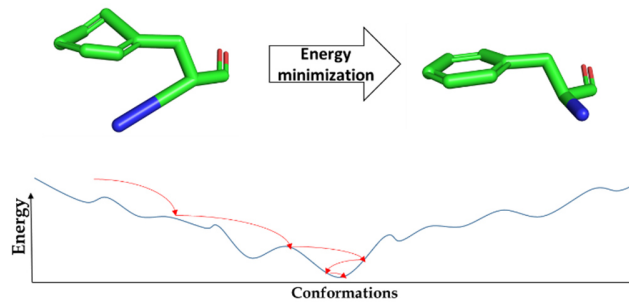
$$V(X) = \frac{1}{2}k_{ij}(r_{ij} - r_0)^2 + \frac{1}{2}k_{ijk}(\theta_{ijk} - \theta_0)^2 + \frac{1}{2}k_{ijkl}[1 + \cos(n\Phi_{ijkl} - \Phi_0)] + \frac{A}{r^{12}} - \frac{B}{r^6} + k \frac{q_1 q_2}{r}$$



Representation of the main ingredients to develop a Molecular Mechanics simulation, (potential energy function, system of atoms and resulting molecular energy landscape) with the example of the PKA-C free energy profile, along the ligand binding and along the open/closed reaction coordinates. Adapted from [107] with permission from the PCCP Owner Societies.(PDB-IDs:1J3H, 1BKX and1ATP)

MOLECULAR MECHANICS

Molecular Mechanics uses minimization algorithms, as the previously explained steepest descent or conjugate gradients minimization methods, to optimize the structure of single molecules, ligand-receptor complexes, or even complex macromolecular systems as molecules immersed in a solvent bulk. It can solve atom misplacements, like atom overlapping, bond length and angles, or unfavorable disposition of partial charges in the molecule, by moving the Cartesian coordinates for each atom, leading to a molecular structure close to a local or global minimum of energy.



Representation of a molecular structure with misplaced geometry and minimization of a bond distance, angle and dihedrals..

The result of molecular energy minimization depends on the starting structure, finding a minimum that in many cases is local rather than the global minimum. Molecular energy landscape is filled with peaks and valleys, and normally a minimization run moves “downhill”. Then, there are no means to explore the overall structural landscape, and with standard methods, there are no means to pass through higher intermediate structures to get to a lower minimum. That is why the initial structure determines the results of the minimization.

Mainly, not big movements in atom position are made, so when looking to minimized macromolecules, the starting structure looks similar to ending structure. Large changes may occur only for significantly distorted structures (stretch bonds).

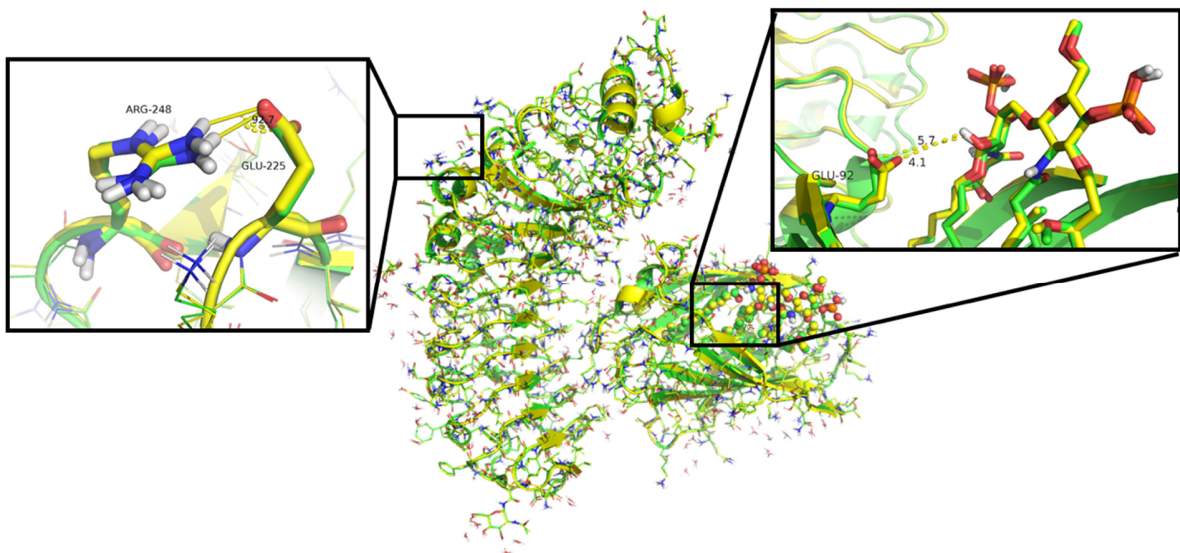


Figure showing the X-ray from the pdb and minimized structure of the 2z65 structure of the TLR4.

MOLECULAR DYNAMICS

Molecular Dynamics (MD) is a method developed to reproduce and understand the dynamical behavior of molecular systems. It is a statistical mechanics approach for simulating complex systems.

MD Simulation will allow studying the dynamical behavior of a molecular system as protein folding or unfolding, binding between molecules, calculating thermodynamic properties such as internal energy, free energy, transitions due to environmental conditions, a tool for sampling the conformations of molecules, structure refinement in conjunction with experimental data, etc.

It is based on performing simulations that move the Cartesian coordinates of each atom by integrating their equations of motion. We can apply the laws of classical mechanics, i.e. Newton's Second law to define a relation between potential energy and the forces in the systems, which allow us to compute the individual forces on each atom from the derivative of the potential energy function with respect to the coordinates. Then knowing that change in position with time gives velocity, and that change in velocity with time, i.e. the acceleration, leading to the forces, we can inversely apply these relations to obtain the increment in the positions of the atoms by integrating the potential energy function.

$$\frac{dU}{dt} = F \text{ and } F = ma \Rightarrow \frac{dU}{dt} = m \frac{dv}{dt} \text{ and } v = \frac{dU}{dt} \Rightarrow \frac{dU}{dt} = m \frac{d^2r}{dt^2}$$

So, if the force F_i exerted on atom i by the other atoms in the system is given by the negative gradient of the potential energy function that in turn depends on the coordinates of all N atoms in the system,

$$F_i(t) = \left(\frac{\partial U}{\partial x_1}, \frac{\partial U}{\partial x_2}, \dots, \frac{\partial U}{\partial x_{3N}} \right)$$

for minor steps (Δt), the following approximation holds, allowing us to compute the position of the atoms in iterative steps:

$$v_i(t + \Delta t/2) = v_i(t - \Delta t/2) + \frac{F_i(t)}{m_i} \Delta t$$
$$r_i(t + \Delta t) = r_i(t) + v_i(t + \Delta t/2) \Delta t$$

Typically, a time step of 1 to 10 fs is used for molecular systems. Thus, a 100 ps (10⁻¹⁰ seconds) molecular dynamics simulation involves 10⁴ to 10⁵ integration steps. Even using the fastest computers only very rapid molecular processes can be simulated at an atomic level.

In order to apply the previous iterative process, we need to initiate MD by assigning initial velocities. We can use the kinetic theory of gases, to assign those velocities in relation with a given temperature. Since the temperature is defined by the average kinetic energy of the system of a molecular system within a box of a certain volume V containing N particles, we can rewrite the definition of the pressure P as,

$$PV = \frac{Nm\overline{v^2}}{3}$$

And combining it with the ideal gas law $PV = Nk_bT$, where k_b is the Boltzmann constant and T the absolute temperature defined by the ideal gas law. With all this, we can obtain the expression,

$$k_bT = \frac{m\overline{v^2}}{3}$$

Leading to an expression for the average kinetic energy of a molecule,

$$\frac{1}{2}m\overline{v^2} = \frac{3}{2}k_bT$$

The kinetic energy of the system is N times that of a molecule, namely $K = \frac{1}{2}Nm\overline{v^2}$. Then the temperature T takes the form

$$T = \frac{m\overline{v^2}}{3k_b}$$

Then, if the system has been energy minimized, that can be understood as freezing the system, the potential energy is zero and temperature is zero. Therefore, to start a MD simulation we need to “heat” system up to desired temperature, by scaling the initial velocities as

$$v = \frac{3k_bT}{2m}$$

With all the previous calculations described before and following subsequent steps or iterations, we can calculate a trajectory in dimensional space of 3N positions and 3N momenta (6N dimensional phase space). Using minor steps in the femto-second range, a MD simulation can have a total duration in the nanosecond-microsecond range.

Note that generalizing these definitions, we can calculate trajectories for all the atoms of a macromolecular system, and it allow us to also represent the water molecules and ions describing the solvent present in the molecular system.

A typical MD simulation protocol will consist in, the initial structure generation, initial energy minimization to obtain a static picture of the frozen system, equilibration of the systems to assign a initial velocities by performing sort MD simulations at increasing temperatures, MD run sampling stable conformations at regular intervals, and final energy minimization of each captured conformation for further analysis. The essential parameters for these MD that are selected by the user are temperature, pressure, time step, dielectric constant, force field, durations of equilibration and MD run, and pH effect by the addition of explicit ions.

Then this approximation allow as to cover a wide range of biological processes as the folding of small peptides that may be in the microsecond range, conformational transitions that may happen in the nanosecond range or collective vibrations as loop and side chain motions involved in the binding between molecules.

Some examples of the application of MD simulation can be found in different fields. It helps on the study of conformational changes or allosteric mechanisms and the study of protein folding. Also, allow applying simulations for sampling of equilibrium ensemble and thermodynamics of flexible systems. MD simulations can also handle modelling and refining of experimental structures obtained by NMR, cryo-electron microscopy and crystallography, prediction of macromolecular structures obtained by docking, homology modelling, etc; or the theoretical study of solvent effects on macromolecular systems the study.

NORMAL MODE ANALYSIS AND ELASTIC NETWORK MODELS

Normal Mode Analysis (NMA) is a molecular modelling technique used to model the vibrations, fluctuations leading to conformational changes of proteins. This technic makes use of a simplified potential energy function defining the system without atomic detail, simplifying each amino acid in the protein structure to one single node centered in its α -carbon.

Each node is related to the others surrounding it, i.e., the amino acids in contact, by a harmonic potential. The nodes in contact are modeled as they would be linked by springs, representing the non-covalent interactions. This definition is called Elastic Network Model (ENM) [108], which is more simplistic in comparison with the atomic detailed molecular force fields, and that is why ENM is classified as a Coarse Grain Model.

The (ENM) provides a simplified representation of the potential energy function of a macromolecular system near equilibrium, where nodes and contacts, i.e., interactions, are represented as a network. Then, each node represents a particle in the three dimensional space, and the edges joining the nodes, act as springs that are represented by a harmonic potential from the equilibrium defined by the experimental structure.

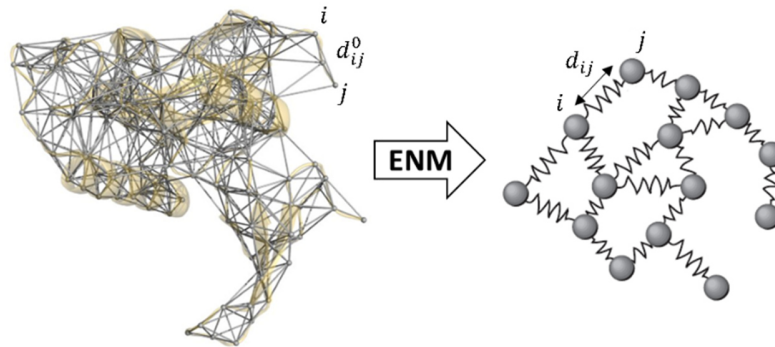


Figure showing the Elastic network model representation of protein Pin1 (PDB ID 3TCZ, ligands removed). Each residue within the structure is represented as a single node at the Ca position, connected to other nodes via Hookean springs. Figure adapted from P. Campitelli et al. [109] and Wikimedia Commons [7] (CC BY).

Then we can define the potential energy of the molecular system by summing up all elements from the this network defined by Elastic Network Model, as springs where we can apply Hooke's law, resulting in the following energy equation,

$$E_{ENM} = \frac{1}{2} \sum K_{ij} (d_{ij} - d_{ij}^0)$$

where K_{ij} is the spring constant describing the different type of interactions (i.e., electrostatic, hydrophobic, etc.), d_{ij} is the actual distance between the nodes i and j during the fluctuation, and d_{ij}^0 the equilibrium distance between the two different nodes defined by the input an experimental structure considered to be stable in the bottom of a harmonic well (i.e., d_{ij}^0 is defined from the experimental structure). The spring constants K_{ij} are the only adjustable parameters in this model, and a variety of methods are used to select their values. Pairwise interactions are predominantly local, and a common practice is to assign a uniform spring constant, $K_{ij} = K$ to all pairs of nodes separated by less than some cutoff distance, and $K_{ij} = 0$ for all others. This definition will correspond to the contact matrix explained in previous sections. It has been found empirically that, when the nodes are taken to be the α -carbons of a protein, a cutoff distance of about 15Å results in residue mean-square fluctuations that correlate well with experimental thermal fluctuations of the atoms in the experimental structure.

Deviations from equilibrium lead to an increment in energy, resulting in a net force towards the lowest energy state. This harmonic approximation to the fluctuations of proteins was first demonstrated to catch the global dynamics of proteins by Tirion in 1996 [110]. Thus, the ENM provides a description of behavior of macromolecules, allowing the study of fluctuations around the equilibrium point with low computational requirements.

NMA could be used with potentials derived from molecular force fields, but those calculations may require initial energy minimization, with risk of distorting the input conformation towards unstable conformations and also incrementing the computational cost.

From the previous definition of the harmonic potential, we can calculate the $3 N \times 3 N$ Hessian matrix of second derivatives respect to the coordinates of the nodes as,

$$H = \begin{pmatrix} \left(\begin{array}{ccc} \frac{\partial^2 E_{ENM}}{\partial x_1 \partial x_1} & \frac{\partial^2 E_{ENM}}{\partial x_1 \partial y_1} & \frac{\partial^2 E_{ENM}}{\partial x_1 \partial z_1} \\ \frac{\partial^2 E_{ENM}}{\partial y_1 \partial x_1} & \frac{\partial^2 E_{ENM}}{\partial y_1 \partial y_1} & \frac{\partial^2 E_{ENM}}{\partial y_1 \partial z_1} \\ \frac{\partial^2 E_{ENM}}{\partial z_1 \partial x_1} & \frac{\partial^2 E_{ENM}}{\partial z_1 \partial y_1} & \frac{\partial^2 E_{ENM}}{\partial z_1 \partial z_1} \end{array} \right) & \dots & \left(\begin{array}{ccc} \frac{\partial^2 E_{ENM}}{\partial x_1 \partial x_n} & \frac{\partial^2 E_{ENM}}{\partial x_1 \partial y_n} & \frac{\partial^2 E_{ENM}}{\partial x_1 \partial z_n} \\ \frac{\partial^2 E_{ENM}}{\partial y_1 \partial x_n} & \frac{\partial^2 E_{ENM}}{\partial y_1 \partial y_n} & \frac{\partial^2 E_{ENM}}{\partial y_1 \partial z_n} \\ \frac{\partial^2 E_{ENM}}{\partial z_1 \partial x_n} & \frac{\partial^2 E_{ENM}}{\partial z_1 \partial y_n} & \frac{\partial^2 E_{ENM}}{\partial z_1 \partial z_n} \end{array} \right) \\ \vdots & \ddots & \vdots \\ \left(\begin{array}{ccc} \frac{\partial^2 E_{ENM}}{\partial x_n \partial x_1} & \frac{\partial^2 E_{ENM}}{\partial x_n \partial y_1} & \frac{\partial^2 E_{ENM}}{\partial x_n \partial z_1} \\ \frac{\partial^2 E_{ENM}}{\partial y_n \partial x_1} & \frac{\partial^2 E_{ENM}}{\partial y_n \partial y_1} & \frac{\partial^2 E_{ENM}}{\partial y_n \partial z_1} \\ \frac{\partial^2 E_{ENM}}{\partial z_n \partial x_1} & \frac{\partial^2 E_{ENM}}{\partial z_n \partial y_1} & \frac{\partial^2 E_{ENM}}{\partial z_n \partial z_1} \end{array} \right) & \dots & \left(\begin{array}{ccc} \frac{\partial^2 E_{ENM}}{\partial x_n \partial x_n} & \frac{\partial^2 E_{ENM}}{\partial x_n \partial y_n} & \frac{\partial^2 E_{ENM}}{\partial x_n \partial z_n} \\ \frac{\partial^2 E_{ENM}}{\partial y_n \partial x_n} & \frac{\partial^2 E_{ENM}}{\partial y_n \partial y_n} & \frac{\partial^2 E_{ENM}}{\partial y_n \partial z_n} \\ \frac{\partial^2 E_{ENM}}{\partial z_n \partial x_n} & \frac{\partial^2 E_{ENM}}{\partial z_n \partial y_n} & \frac{\partial^2 E_{ENM}}{\partial z_n \partial z_n} \end{array} \right) \end{pmatrix}$$

After defining the harmonic potential of the Elastic Network Model, and the Hessian matrix derived from the potential, we can define the Normal Mode Analysis (NMA) as the diagonalization of H yielding to the eigenvector (v_m) and eigenvalues (λ_m) decompositions,

$$Hv_m = \lambda_m v_m \Leftrightarrow (H - \lambda_m I)v_m = 0$$

Eigenvectors, i.e. Normal Modes, define the direction of a certain fluctuations and, and eigenvalues are related with the frequencies of the fluctuation (ω_m) in the direction of the normal mode by,

$$\omega_m = \sqrt{\lambda_m}$$

When computing Normal Mode Analysis the 6 first modes have zero eigenvalue and correspond to rigid-body rotations and translations of the system. The remaining $3N - 6$ Normal Modes are studied as the possible fluctuations of the molecular system, having 3-vector component for every node.

It is important to note that the lowest frequency Normal Modes, are the ones to be considered as biologically relevant as they may capture the slow collective motions corresponding with conformational changes.

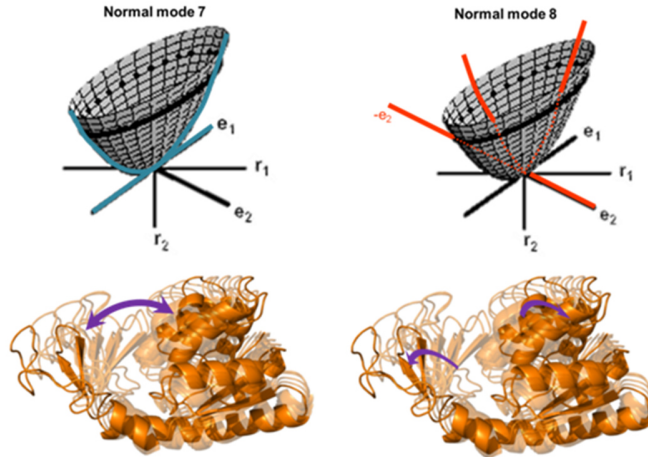


Figure showing the two lowest frequency Normal Modes from the Elastic Network Model defined by the leucine binding protein in the absence of ligand (pdb-id 1usg). Note that the lowest frequency Normal Mode captures the conformational change observed between ligand free and ligand bound conformation.

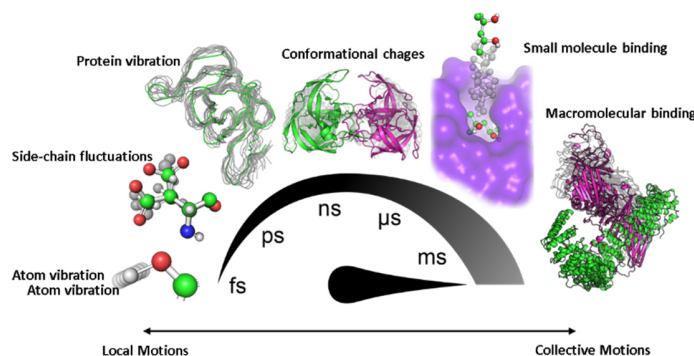
This approximation is usually applied to proteins, but it can be extended to larger systems using different definitions of the nodes of the ENM for even macromolecular complexes. For example, W. Hu *et al.* [111] studied the intrinsic dynamics of a DNA octahedron using different types of Elastic Network Models (ENMs), defining nodes as “DNA nanocages”, describing the intrinsic flexibility of DNA double-helices and hinges. This was done through the calculation of the square fluctuations, as well as the intrinsic collective dynamics in terms of cross-collective map calculation coupled with global motions analysis with NMA.



Representation of Elastic Network Model applied in molecular simulations, the intrinsic dynamics of a DNA octahedron. Image from W. Hu *et al.*[111] (CC BY 4.0).

APPLICATIONS OF NORMAL MODE ANALYSIS

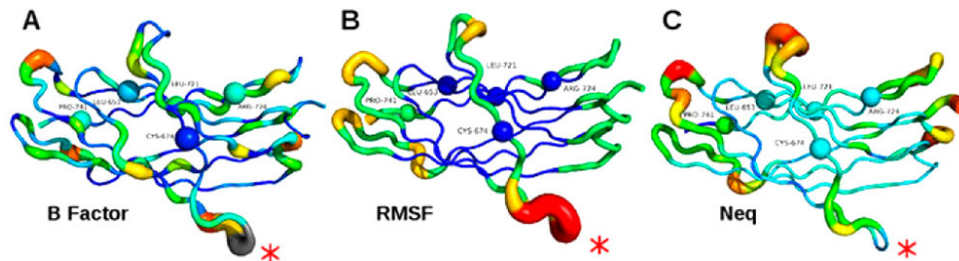
The frequency spectrum of protein conformational changes can be studied respect to the collectivity of that specific motion, i.e. the number of atoms that displaces a certain conformational change. When studying collectivity versus frequency over a large set of know protein movements, it was found that half of the protein systems could be modelled by looking at the two lowest frequency normal modes [112]. Then computing the Normal Modes from an Elastic Network Model allow measuring the deformation of each residue in the molecule, defining all possible ways where the molecule can deform are measured. This gives a measure of how flexible a protein can be.



Hierarchy of principal motions in protein dynamics. From left to right: bond vibrations (fs–ps), side-chain rotations (ps–ns), backbone fluctuations (ns), loop motion/gating (ns–ms), ligand binding/unbinding events (>100 ns), and collective domain movement (>μs). Figure adapted from B. Surpeta *et al.* [113] (CC BY 4.0).

The thermal factors or B-factor of an experimental crystallographic structure describe the displacement of the atomic positions from an average value provided in the structure, and they are computed in all crystallographic structures. When the regions of the protein are flexible, the larger the displacement from the mean position given by the experimentalist will be. Protein B-factors are found in the last column of the PDB file and means. Well defined regions have low B-factors (blue/green) Poorly defined/more mobile regions have high B-factors (yellow/orange/red, right figure). From Normal Modes, we can estimate the B-factors based on the deformability

values, and their values can be compared to check if there is correlation between the B-factors from the normal modes and the B-factors that we observe in the experiments.

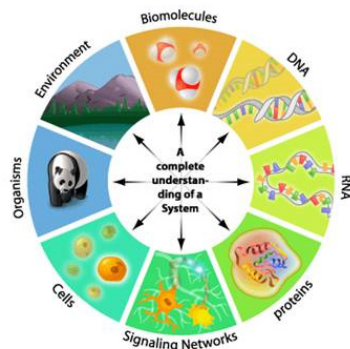


Protein flexibility of Calf-I: (A) B-factor values, (B) RMSF values, and (C) Neq values. Local structure ranked from rigid (thin blue line) to flexible (thick red line). Missing atoms in grey, B-factor cartoon. Figure from M Goguuet, et al. [114] (CC BY-NC-ND 4.0).

If the experimental structure of a protein structure is known in two different conformations, Normal Mode Analysis also allows comparing two conformations and may be used to determine the contribution of each normal mode to its conformational changes. This can be done, comparing the directions of the Normal Modes with the displacement vectors of the α -carbons from the open and closed (bound and unbound) conformation. If there is an overlap between those displacement vectors and any of the low frequency normal modes, it means that the conformational change is well described by NMA, and it is possible that motions observed between the bound and unbound conformations are due to the internal fluctuations of the molecular system. This implies that this conformational change is better described by conformational selection model rather than induced fit.

INTRODUCTION TO DATABASES

The goal of this section is to understand the importance of databases in modern molecular biology, putting special emphasis in its biomedical applications. In this section, we will show some of the most common databases, using these as example to understand the wide array of information that has been stored and the great possibilities that this offers to solve real biomedical problems. This section offers an overview of the content of these databases, as well as the importance of related tools, usually linked to the same database, that allow us to search and analyze the data in order to obtain a biological interpretation of the molecular mechanisms occurring at the cellular level. Finally, by reviewing these databases, we will provide a better understanding of the importance of the relationship between all the biological systems.

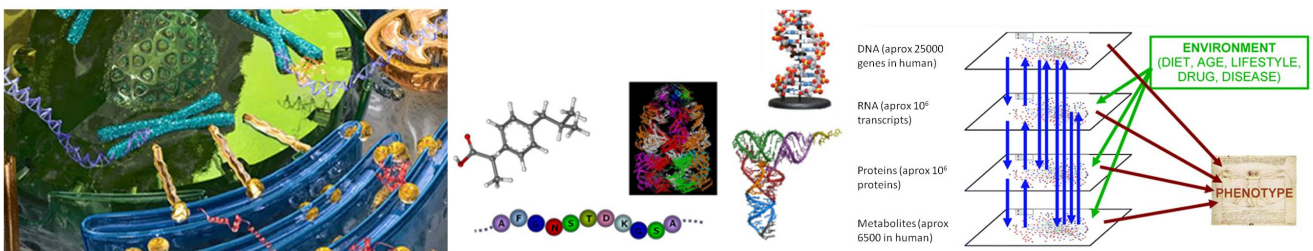


Most importantly, BROWSE and know where and how to find relevant biological information

Schematic representation of the goal of this section

INTRODUCTION

The NIH defines biological systems as a group of biological molecules that interact in an established pattern [115]. These molecules have a wide variety of properties, structures and functions. They can be complex polymers, such as polynucleotides or proteins, with different structures, sequence variations, expression levels, or a wide array of other smaller molecules with various physicochemical properties, involved in several pathways or chemical reactions. Furthermore, these molecules interact between them in a myriad of possibilities, and these interactions are fundamental to perform a determined function or biological process, and therefore to understand the distinct phenotypes. Many years ago, doctors had to base their diagnosis in a small number of phenotypic observations. Nowadays, with the improvement of technological instruments that allow us to measure a great number of biological data simultaneously, the basis for diagnosing a certain condition is supported with more precise and better observations. Also, the analysis and interpretation of these interactions will allow to build predictive mathematical models that can provide a more personalized and detailed interpretation of the outcomes of a perturbation in the molecular mechanisms, such as for example the treatment with a drug, or the knock out of a gene.

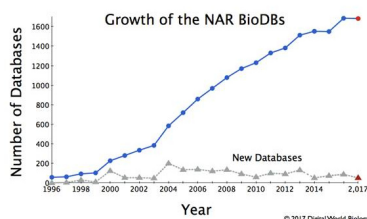


Different types and interactions between molecules

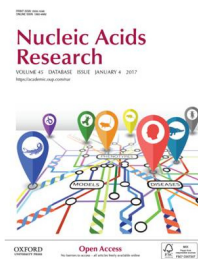
WHAT ARE (BIOLOGICAL) DATABASES?

A database is an organized collection of data. It needs to be organized so that we can work with it, to avoid redundancies, easy search, sharing of data, etc. Databases are integrated in our day-to-day work, not only in research, but also in the industry or communication systems. Most of our daily resources and activities generate data. You don't need to be an expert to consult a database, because most of them have easy to use graphical user interfaces (GUIs).

Concerning molecular and cellular biology, the journal *Nucleic Acids Research* (NAR), publish a yearly issue entirely dedicated to advances, changes or new biological databases [116]. The number of these databases have been constantly growing since the early 90s and is nowadays more than 1600.



http://oxfordjournals.org/our_journals/nar/database/a



The 27th annual *Nucleic Acids Research* database issue and molecular biology database collection

Daniel J Rigden ✉, Xosé M Fernández

Nucleic Acids Research, Volume 48, Issue D1, 08 January 2020, Pages D1–D8,

<https://doi.org/10.1093/nar/gkz1161>

Published: 22 December 2019

NAR Database issue, number of 2020. Adapted from D. J. Rigden et al. [116]

The NAR journal classifies the biological databases in 15 different categories, including protein and nucleotide sequences, structural databases, metabolic and signaling or human and vertebrate databases. The most usual terms in the name of the more than 1600 databases (apart from database per se) are protein, followed by gene, genome, human, sequence, and data [116]. In addition, structure, genomic, interaction and expression are widely repeated. While data from DNA sequencing seems to be the most interesting and widely used, to analyze how genotypes impact phenotypes requires to understand the relationship between sequence, structure and function. It is therefore interesting to note that the most common terms describing biological databases would include words that describe this relationship. It is also interesting to note the great number of unique words, showing that many databases are focused in small pieces of information, interesting to only one or a few research groups. Some examples include databases related to waterfleas, mites, honey, plexipus or bananas. In this section, because the rest of the book is dedicated to structural data, we are going to focus in functional data.



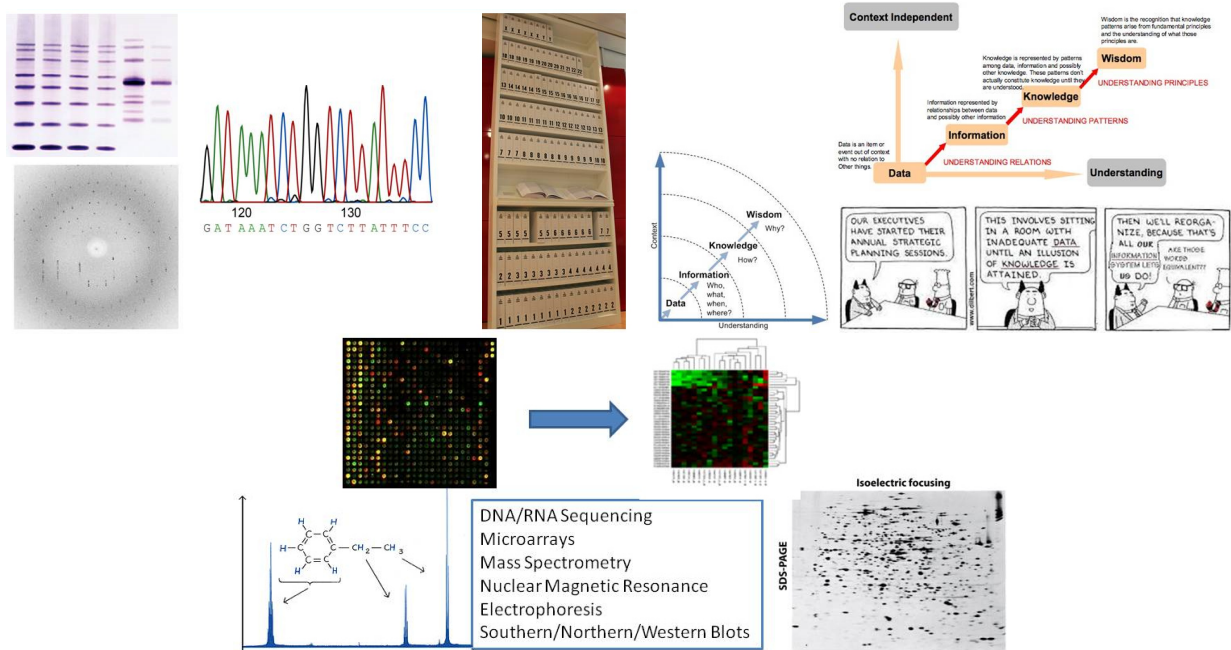
Categories of databases according to NAR special issue. Adapted from D. J. Rigden et al. [116]

WHAT IS (BIOLOGICAL) DATA?

Data is the result of a measurement. Biological data is a particular type of data, generated in the context of a scientific research. Measurements in scientific research are generally nontrivial, but mostly the result of a complex workflow using very diverse methodologies (ex. sequencing nucleotides, or the band corresponding to a protein in a western Blot).

There is sometimes misunderstanding between data, information and knowledge. Data is a fact, is objective and non-abstract. However, information and knowledge are subjective and require a mental process. Biological data organization in databases is useful to facilitate this mental process. In general, we can have primary databases, that store pure biological data, and secondary databases, that store not only objective data, but also information obtained from the relationship of different pieces of data between them.

Therefore, we have raw data, like the isoelectric point of a protein, which is extracted directly from the experimental technique or we can also have stored processed data. In fact, most of the times, we need to process the data, to interpret them, and obtain information and knowledge [101]. For example, and mainly in modern analytical tools, information is not evident in the data (like in the file containing the expression profile of an amount of genes from an organism). Data processing is done in different sequential steps, depending on the technique and the information required.

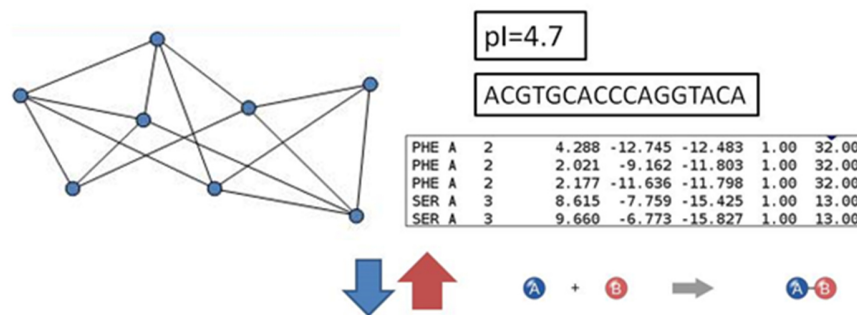


Types of biological data. Adapted from N. J. Lozano et al. [101]

CHARACTERISTICS AND TYPES OF BIOLOGICAL DATA

Despite its diversity, biological data has some common characteristics that explain the way they are commonly stored in databases [117]. In general, biological data is:

- Heterogeneous, even in the same category, biological data has many appearances, from atoms to population studies, from nucleotides to protein structure, including interactions, cells, phenotypes or physiological studies. The storage of this wide variety of data explains the number of different databases and the grand number of links between them.
- Complex, biological data is not easy to obtain or interpret (for example isolation of a protein). In general, biological experiments consist on a great number of steps, which increases the probabilities of making mistakes. As said before, the interpretation of this data is not always trivial.
- Qualitative or quantitative, data can be qualitative, like the function of a protein, or quantitative, like a molecular weight. Also, information in biological databases can be stored in many ways, that may seem to be redundant, as the function of the protein can be defined in many different ways.
- Dynamic, data may change with new discoveries, or interpreted in different ways in the future thanks to the improvement of technology (for example, since its discovery, the sequence of insulin has been reviewed more than 120 times). This is important as databases are in constant change.
- Data can be experimental, computational or it can come from indirect interpretations of experiments.



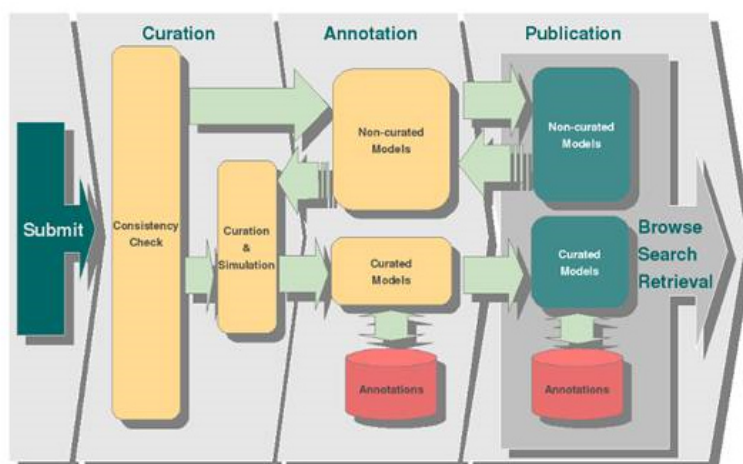
Characteristics of Biological Data. Adapted from et al. M. Chagoyen [117]

Due to this complexity, data can be stored in many different types. Some of the most common ones are scalar (as for example the isoelectric point of a protein), chain (like the sequence of a nucleotide), vectorial (like the dynamics of a biochemical reaction), graph (especially interesting in phylogenetics or protein interactions), temporal (like the changes of expression of a gene in different times) and spatial (like the structural coordinates of atoms in a molecule).

DATA QUALITY, REVISION, ANNOTATION AND REDUNDANCIES

Data quality depends totally on the curation process. Curation can be manual or computational. It consists in the review by computers or human researchers (curators) of the scientific soundness of the data, checking the experimental procedure and the reliability of the data obtained. Databases can be curated or non-curated. Revision of data is the analysis and integration of the stored biological information in order to add additional value through annotation and integration of data. It is essentially the process to give benefit to the data, and obtain information and knowledge [118]. There are so many databases that we need cross references to link information from one to another. Curators are also in charge of this revision and cross-referencing process. Within all this great amount of data and databases, there is obviously a great deal of redundant data, either interDB (data repeated over different databases) or even intraDB (data repeated within the same database). These can happen because either the experiment is repeated and actualized, or because another group did the same experiment, and uploaded the

information in the database separately (and was not detected by the computerized or manual curators). Redundant data can even be inconsistent sometimes, which causes problems in interpretation and reliability of studies, and it also makes a database to need more storage capacity. Curators try all the time to eliminate redundant and inconsistent data.



Data Quality, revisions and annotation. Adapted from S. Burge et al. M. Chagoyen [118]

WHY DATABASES ARE ESSENTIAL IN MODERN MOLECULAR BIOLOGY

The sum of the complexity of biological systems and their interactions, together with the huge amount of data generated that needs to be processed in an exhaustive way, highlights the importance of the generation of databases and computer tools for the analysis of biological information [119]. The amount of high throughput technological developments produced in the last decades (nucleotide sequencing, spectrometry technologies, etc.) have widely increased the pace at which data is being gathered. The next challenge is to make the analysis, access and management of data as efficient as the generation [120].

Databases also help in the reutilization or reevaluation of the data to obtain new information that was not noticed before, or that was not possible to interpret in the past due to lack of technology. They also allow sharing the data or information between different groups, therefore increasing the rate at which relevant interpretation of the data can be obtained. In order to do this, data should be accessible online and easy to download, as it is in most of the main databases (although the importance of personal privacy in data should be considered thoroughly in some cases). There is a global awareness in research groups to share data, and it is normally a requisite for publishing to upload data in online and free repositories or databases.

Along with databases, computer tools are essential to obtain information from data. The wide amount of data needs these computer tools to obtain relevant information. Databases and web servers are the best places to gather these tools, and provide additional commodities to users in this regard.

Also, some databases are very general while some are very specialized. For best results, we often need to access multiple databases.

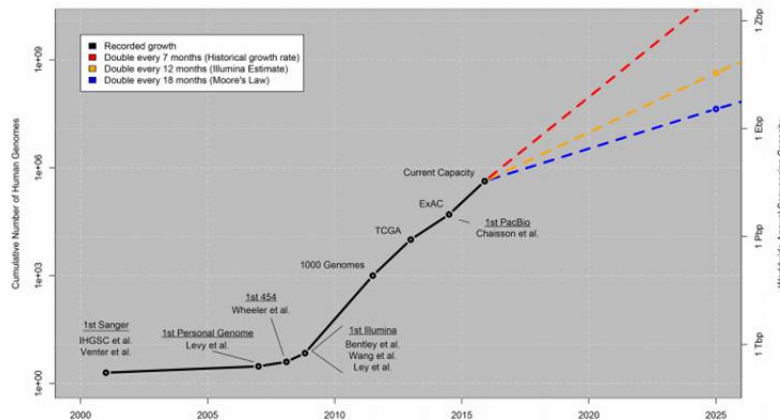
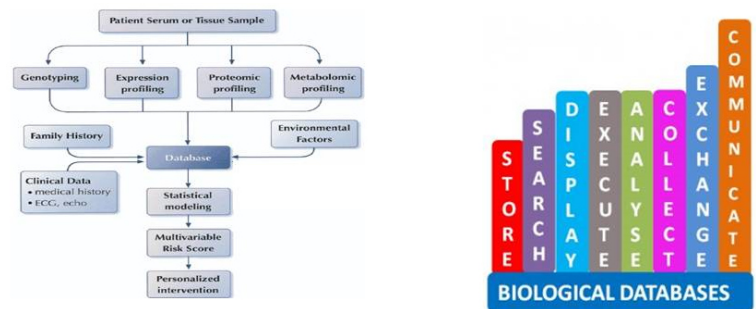


Table 1. Four domains of Big Data in 2025. In each of the four domains, the projected annual storage and computing needs are presented across the data lifecycle.

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

Growth of DNA Sequencing

Importance of databases and data in biology. Adapted from Z.D. Stephens et al. [121] (CC BY 3.0).

Therefore, we summarize the requisites for a good database. First, accessibility, where it is mandatory that data is freely available to all the scientific community. It is also important for databases to be online. Second, data should be complete and be actualized periodically, whereas all data and cross references whenever available should be included, and due to the dynamic nature of biological data, databases should be actualized. Third, databases should have intuitive and friendly interfaces, that allow to incorporate and download data and tools to analyze them. Forth, quality where data should be curated both manually and automatically.

So databases require a lot of effort for design, implementation, maintenance, organization, annotation, supervision, deposit and storage. Therefore, the most commonly used databases are maintained by big consortiums and collaborations between institutions and research groups, as well as fully dedicated curators.

INTERNATIONAL CONSORTIUMS

The amount of work required for a database requires big International Consortiums to support them. The two main ones are integrated in the EMBL (European Molecular Biology Lab) and the NCBI (National Center for Biotechnology Information). Due to the great number of databases, each Consortium has built integrated interfaces with many DBs incorporated and with common search engines (Entrez).

An example is the EBI (European Bioinformatics Institute): is part of the EMBL. It started the year 1992 in England to store the great amount of sequences from the Sanger Institute. Actually supports Ensembl (for DNA) and Uniprot (for proteins).

The NCBI, part of NIH is in Maryland. Founded in 1988 to develop information systems in the field of molecular

biology. It supports Genbank and PubMed.

Anybody can incorporate data in one of these databases, through the web interface, with an institution ID. It will later be checked by researchers in each Consortium. Automatic and manual curation processes are later performed by researchers in the Consortia looking through research papers and other databases. They will check for nomenclature, redundancies, etc.



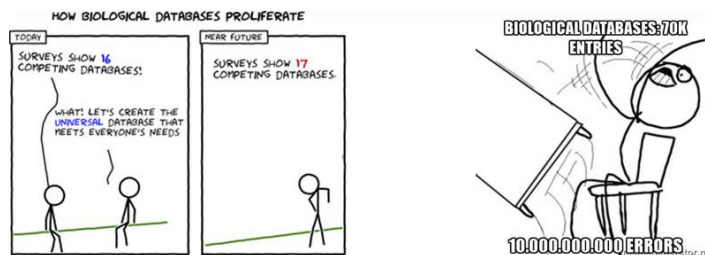
International Consortia responsible of molecular databases

POTENTIAL PITFALLS IN DATABASES

As said before in different sections, the problems with using biological databases include incomplete information, data spread over multiple databases, redundant information, various errors, sometimes incorrect links, constant change, nomenclature, low quality of data and inconsistencies [122].

Databases contain mistakes (although in general these mistakes are low as a proportion of total data). The problem is that these mistakes are usually difficult to correct

Therefore, when using a database, you have to use information intelligently, always ask yourself if the conclusions make biological sense and if you may require further analyses or experimentation.

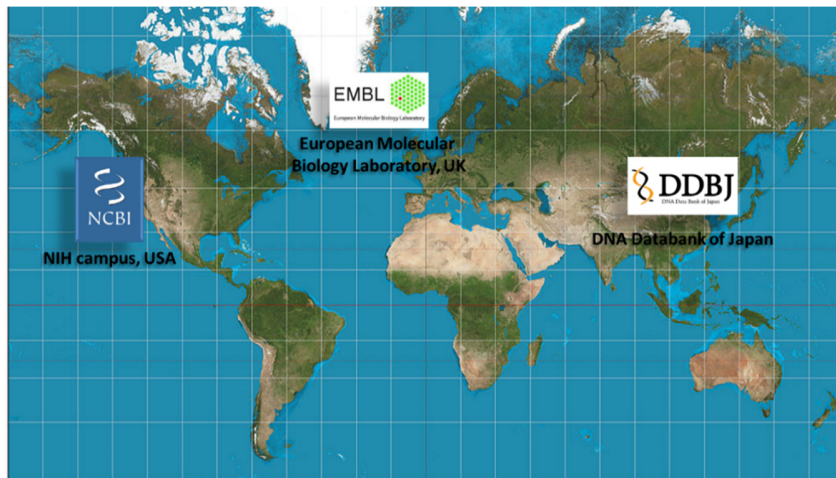


Pitfalls in Molecular Databases

NUCLEOTIDES DATABASES

These databases gather mainly sequence data, but there are also secondary databases with annotated data about genes or regulatory regions. Since the early nineties, the amount of DNA sequences from different organisms has been constantly growing at an exponential rate. Both the NCBI and the EMBL have databases for storing all these sequences, namely, GenBank from the NCBI and the ENA (European Nucleotide Archive) from the EMBL. Together with the DNA DataBank from Japan (DDBJ), they have formed a big Collaboration (International Nucleotide Sequence Database Collaboration) that exchange sequences on a daily basis. These three databases store and classify data depending on taxonomic divisions or in the sequencing strategy.

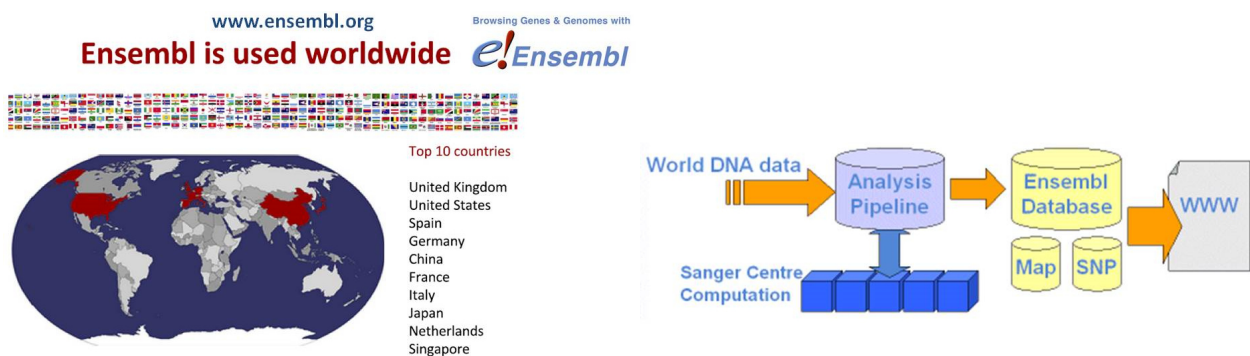
In order to store information and annotated data, the EMBL lab has built a secondary database, called Ensembl, that is going to be used as an example of a curated, secondary nucleotide database [123].



Most extended Nucleic Acid Databases represented in the world map.

ENSEMBL NUCLEOTIDE DATABASE

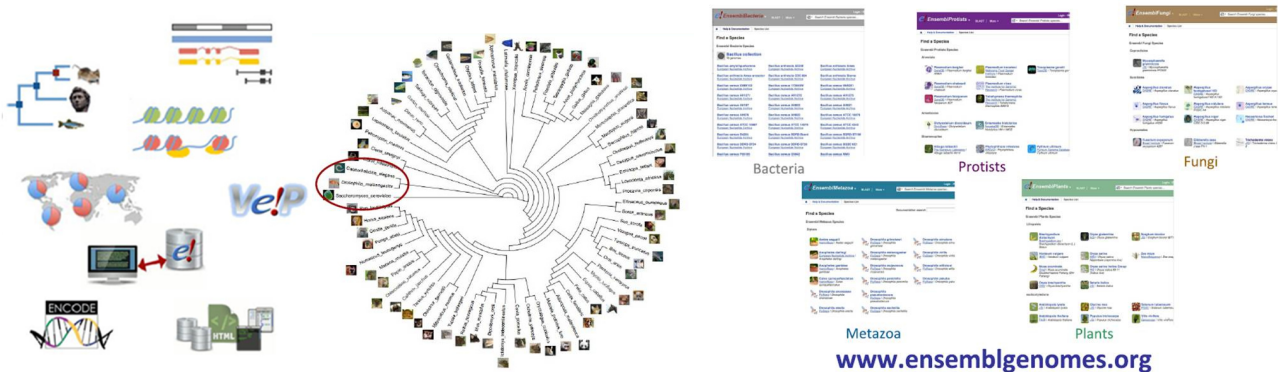
Ensembl is a joint project between EMBL European Bioinformatics Institute and the Sanger Institute to develop a software system, that produces and maintains automatic annotation on eukaryotic genomes. Ensembl is primarily funded by the Wellcome Trust. Ensembl is used worldwide, mostly in USA and Europe. Ensembl utilizes raw DNA sequence data from public sources (Genbank, ENA and DDBJ) and creates a tracking database (The “Ensembl database”). With computational and manual approaches, Ensembl joins the sequences - based on a sequence scaffold called “Golden Path”, and automatically finds genes and other features of the sequence. It then associates sequence and features with data from other sources and provides these information in a publicly accessible web based interface (www.ensembl.org) [123]



Ensembl Data and Annotation. Adapted from Ensembl Genome Browser <https://www.ensembl.org> [119]

Ensembl has data and information regarding comparative genomics, evolution, sequence variation, disease data and transcriptional regulation for example. Ensembl annotate genes and has incorporated tools that can compute multiple alignments (through a BLAST tool), predict regulatory function or the effect of a list of variations (Variant Effect Predictor –VEP-) for all supported species.

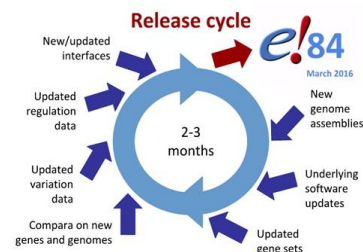
Ensembl is a gene and transcript database of vertebrates (it actually has information on more than 70). It also incorporates 3 non vertebrate species (yeast, worm and fly) for control (to test new tools or updates), research (most widely used species in research labs) and historic purposes (were among the first ones to be sequenced). The data and information on non-vertebrate organisms is found in a different web, called EnsemblGenomes (www.ensemblgenomes.org) that is subdivided in five categories (Ensembl Bacteria, Ensembl Fungi, Ensembl Metazoa, Ensembl Plants and Ensembl Protists).



Images from Ensembl Database. Adapted from Ensembl Genome Browser <https://www.ensembl.org> [119]

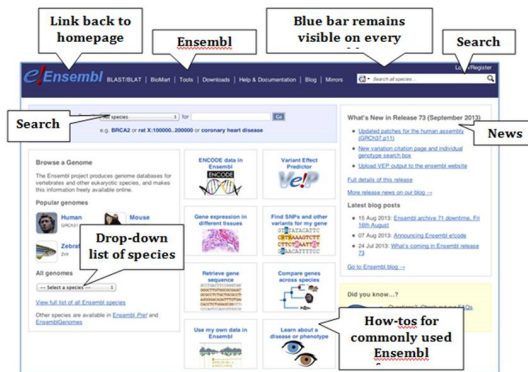
The main difference between Ensembl and EnsemblGenomes is that the first one is annotated by the Ensembl Consortium while the latter is, due to its bigger magnitude, annotated through collaborations between the scientific communities. The annotation release cycle in Ensembl lasts approximately 3 months. Each release begins with the incorporation of new genome assemblies, and is followed by updates in software, gene sets, variation and regulation data, tools and the interface.

	Ensembl	EnsemblGenomes
Released	2000	2009
Species	Vertebrates (fly, worm and yeast as outgroups)	Non-vertebrates (protists, plants, fungi, metazoa, bacteria)
Annotation	by Ensembl	in collaboration with the scientific communities
URL	www.ensembl.org	www.ensemblgenomes.org



Ensembl cycle release. Adapted from Ensembl Genome Browser <https://www.ensembl.org> [119]

The information that can be found in Ensembl is both global (species specific information, i.e. karyotype, phylogenetic trees, number of genes in a specie) or individual information about genes and regulatory regions for each organism. For example, for a particular gene we can find information on the name, organism, revision status, summary info. In addition, we can find the gene context (location and sequence accession), information regarding the different transcripts and gene products (sequence, transcripts, protein sequence). All this information is supported with links to bibliography and cross references to other databases, as for example OMIM [124] for diseases or dbSNP [125] for genetic variations. The Ensembl interface also shows functional information regarding the related phenotypes, pathways and ontologies corresponding to each gene.



Ensembl web page description. Adapted from Ensembl Genome Browser <https://www.ensembl.org> [119]

Finally, Ensembl provides a location tab, where all this information can be overlaid with the sequence position. In this case, Ensembl shows a color and number code for data that has been automatically or manually annotated as well as for coding genes, pseudogenes, etc. (view Figure below). Havana (Human and Vertebrate Analysis and Annotation) is a gold standard both manually and automatically annotated reference genome. It is coded with the color yellow and it is the highest confidence sequence and annotation available in Ensembl.

Automatic annotation*	Manual annotation*
<ul style="list-style-type: none"> many species genome-wide at once ~ 4 months 	<ul style="list-style-type: none"> fewer species gene by gene many years

* based on experimental, biological evidence (INSDC, UniProtKB...)

Gene models in Ensembl

Legend:

- protein coding
- merged Ensembl/Havana
- pseudogene
- processed transcript
- RNA gene
- RNASeq gene

Automatic Manual

Goal: Generate set of well-supported genes

Alternatively spliced transcripts

rich and comprehensive annotation

Gold (identical annotation) = Automatic + Manual

Ensembl web page screenshots. Adapted from Ensembl Genome Browser <https://www.ensembl.org> [119].

TOOLS IN NUCLEOTIDE DATABASES

BLAST

As commented in previous sections, BLAST (Basic Local Alignment Search Tool) is a tool developed by the NCBI that helps us search between similar sequences [126]. It is usually incorporated in different nucleotide or protein databases. It allows the comparison between nucleotide or protein sequences, and therefore can align a query sequence to a reference sequence in a database. It also calculates the statistical significance of matches giving a score to each of them. It is like Google but for nucleotide sequences.

Some of the applications of this BLAST tool can be the evaluation of evolutionary relationships, identification and annotation of sequences, diagnosis of diseases, *in silico* evaluation of primers, or to model genomic structure. As an example, Darwin's comparison of morphological features of the Galapagos finches led him to postulate the theory of natural selection. With BLAST, when you compare the sequences of genes and proteins, you are performing the same type of analysis, just at another level.

There are different scoring systems, penalizing more or less the gaps, the mismatches and the misalignments. In general, the highest score has more probabilities of being the correct match for our query sequence. However, other parameters, such as the length coverage (query start to query end) of the alignment, the percentage of identity or the E-value (probability that the alignment is due to random chance) have to be taken into account.

Align:
THIS IS A RATHER LONGER SENTENCE THAN THE NEXT
THIS IS A SHORT SENTENCE

THIS IS A RATHER LONGER - SENTENCE THAN THE NEXT
|||| | | --*|--- |---| - ||||| |||| - - - - -
THIS IS A --SH-- -O--R T SENTENCE - - - - -
OR
THIS IS A RATHER LONGER SENTENCE THAN THE NEXT
|||| | | - - - - - ||||| |||| - - - - -
THIS IS A SHORT- - - - - SENTENCE - - - - -

- Match score: +1
- Mismatch score: +0
- Gap penalty: -1

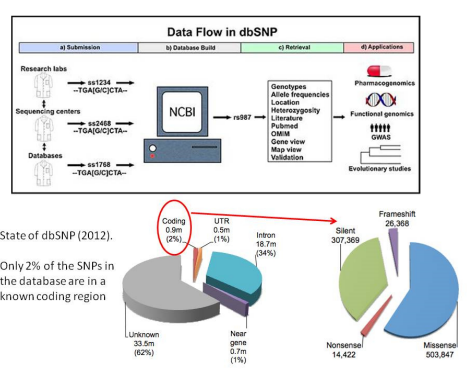
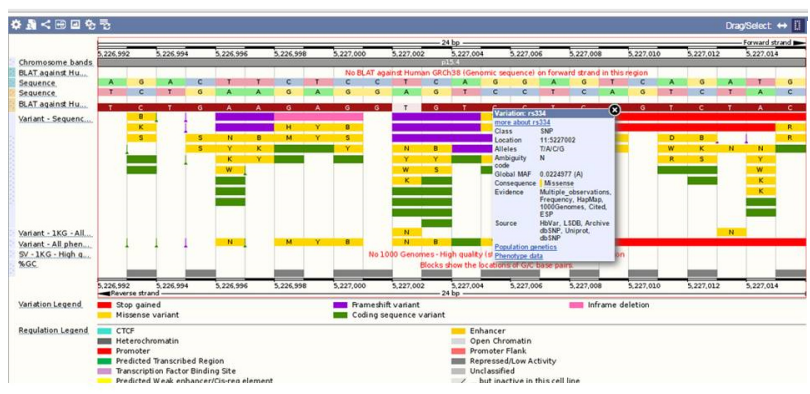
ACGTCTGATACGCCGTATAGTCTATCT
 ||||| |||| | | ||||| |||||
-----CTGATTCGC--ATCGTCTATCT

- Matches: $18 \times (+1)$
- Mismatches: 2×0
- Gaps: $7 \times (-1)$

Score = +11

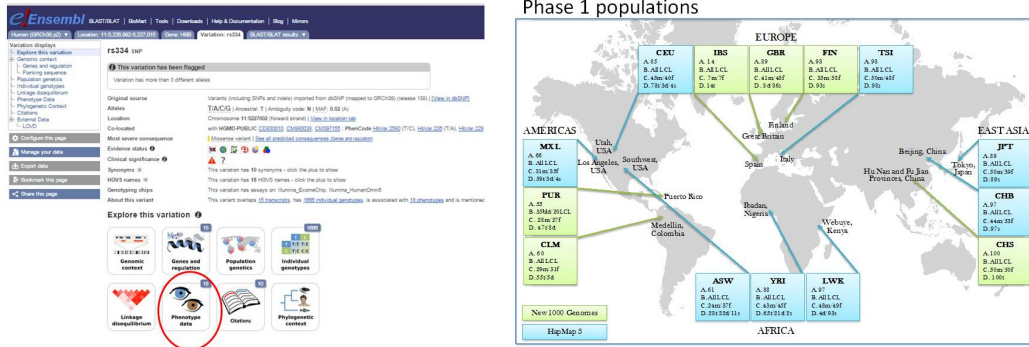
Representation of what a sequence alignment is with the representation of the alignment score. Adapted from J. Chang [127].

Ensembl has a BLAST tool incorporated. One of the great advantages of the integration of this tool in databases is that once the query has been aligned, you can overlay data from different resources on top of the query and find relevant and additional information regarding your search. For example, if you find a mismatch in a query sequence, you can compare it with variant information from different databases, like the dbSNP, and discover if that mismatch corresponds to a known variation. If that is the case, you can find information in dbSNP regarding known phenotypes, poblational information, allele frequencies etc.



Information from dbSNP [125].

The information in the dbSNP or in Ensembl can be complemented with the results obtained from the 1000Genome Project, a multi collaboration project where the human DNA sequence of different races was obtained in order to build a human reference genome with information of different alleles and variation frequencies. The 1000 Genome Project is an international project to construct a foundational data set for human genetics that may discover virtually all common human variations by investigating many genomes at the base pair level [128]. It is a Consortium with multiple centers, platforms, phases and funders.



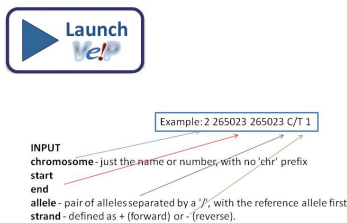
Information from OMIM and 1000 genome Project [128].

VARIANT EFFECT PREDICTOR

The Variant Effect Predictor tool (VEP) determines the effect of a list of different variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions [129]. The input is a file with the coordinates of the variations and the nucleotide changes in each of them. The output of the VEP analysis is a table with: genes and transcripts that are affected by the variants; the location of each variant (coding sequence, non-coding RNA, up or downstream). the consequence of the variation in the protein sequence (missense, frameshift mutation, stop, intronic...); links to databases of known variations such as dbSNP; the allele frequencies from the 1000 Genome Project and the results from the score prediction of changes in the protein sequence (SIFT and Polyphen).

The SIFT predicts whether an amino acid substitution is likely to affect protein function based on sequence homology and the physico-chemical similarity between the alternate amino acids [129]. Polyphen is another prediction algorithm that predicts the effect of an amino acid substitution on the structure and function of a protein using sequence homology, Pfam annotations, 3D structures from PDB where available, and a number of other databases and tools (including DSSP, ncoils etc.) [130].

Variant Effect Predictor: VEP



Tool	Numerical Value	Qualitative Prediction
PolyPhen	0	Benign
PolyPhen	0.8	Possibly damaging
PolyPhen	1	Probably damaging
SIFT	0	"deleterious"
SIFT	1	"tolerated"

Variant Effect predictor tool [129].

PROTEIN DATABASES

Proteins have a wide amount of biological complexity. Their physicochemical properties, as well as structure and function depend on the amino acid sequence, specially depend on the lateral residues in the chain, but also in the peptide backbone. There are 20 possible coding amino acids, and there are other non-coding possibilities. Due to this complexity, there are a number of different possible classifications of protein databases. In general, we will deal with sequence databases, structural databases and interaction databases.

The structural databases, such as Protein Data Bank (PDB) have already been treated in other sections of this course. Therefore, in this case we are going to focus on the functional (sequence) and the protein-protein interaction databases. The server that gathers the biggest number of protein databases and information belongs to the Swiss Institute of Bioinformatics and is called ExPASy (Expert Protein Analysis System). The most common functional database for proteins is UniProt [131].



Types of protein databases.

UNIPROT

UniProt is a joint database from the PIR (Protein Information Resources), the EMBL and the Swiss Institute. The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information. More than 95% of the protein sequences provided by UniProtKB come from the translations of coding sequences (CDS) submitted to the ENA, GenBank or DDBJ nucleotide sequence resources. UniProt makes use of a subset of evidence codes from the Evidence Code Ontology (ECO) to indicate data origin [131]. UniProt, as in Ensembl, has primary sequence data from proteins (stored under the UniRef and UniParc archives) and annotated functional information of proteins from different organisms.

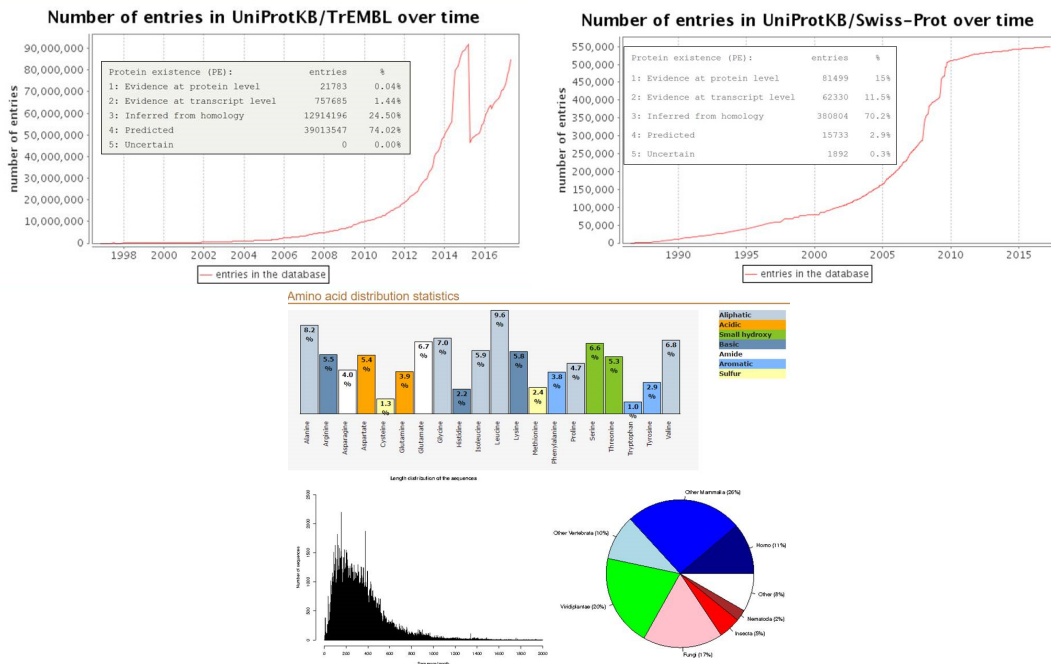
The annotation in UniProt consists on the description of the following items: function of the protein; post translational modifications (PTMs) such as ubiquitination, glycosylation, phosphorylation; protein domains and sites, like homeoboxes, zinc fingers, catalytic pockets; structural information (primary sequence, secondary structure and links to 3D tertiary and quaternary structure); similarities and alignments to another proteins and proteins in other organisms; diseases and phenotypes associated with variants in the protein; subcellular location; interactions with other proteins.



UniProt Knowledgebase [131].

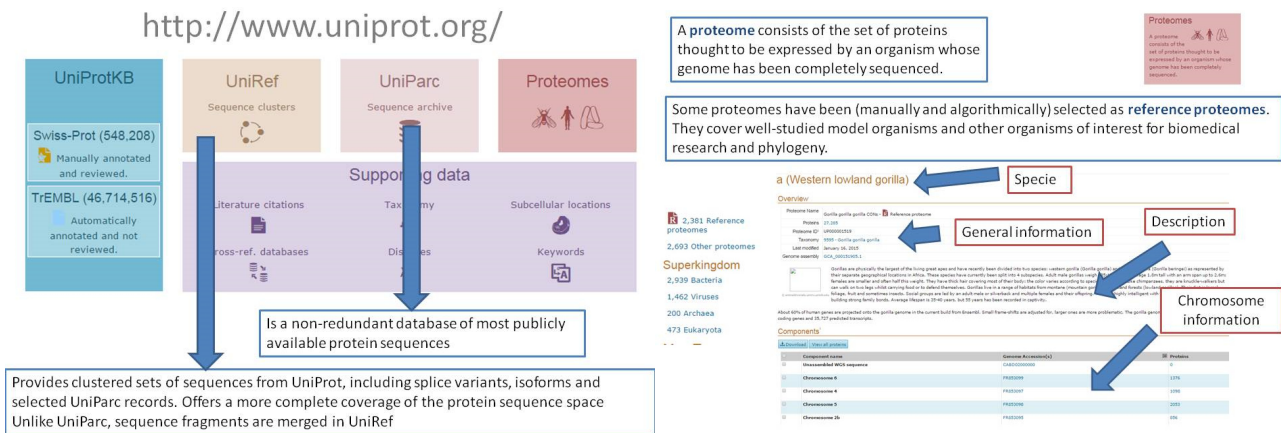
The Uniprot Knowledgebase (UniprotKB) is divided in two main categories. Firstly, the TrEMBL (transcription EMBL) contains more than 80 million proteins from different organisms that have been obtained from computational translation of DNA sequences. These proteins are therefore automatically annotated and are not reviewed. The other category in UniprotKB is called Swiss-Prot and has about 500000 proteins that have been annotated and manually reviewed and curated. These ones are marked with a golden symbol and the information is therefore more reliable.

Uniprot has global statistics on the number of amino acids, polarity, protein size, that can be seen in the figures below.



Statistics from Uniprot Database [131].

As in Ensembl, Uniprot gathers information on the complete set of proteins in different organisms (proteome) as well as information on individual proteins. In this case, it has collected information for thousands of organisms, not only vertebrates. There is a number of reference proteomes (marked with a red symbol) that are very well studied, annotated and curated organisms of interest for biomedical research and phylogeny.



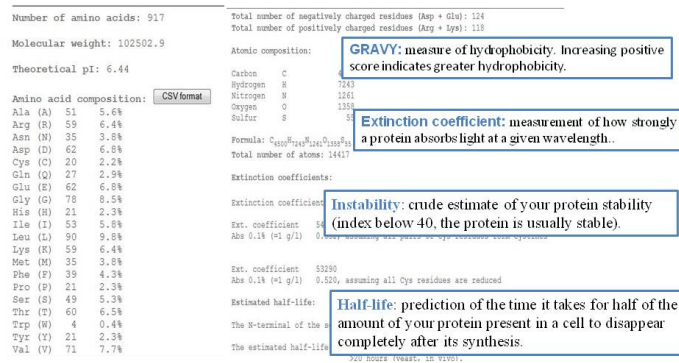
Data storage and information from Uniprot [131].

UNIPROT TOOLS

There are many tools that can be used. For the purpose of this section, and to reinforce the contents of other sections, we are going to explain two tools, useful to obtain predictions of physicochemical parameters of the proteins based on their amino acid sequence. These physicochemical parameters can be related to the 3D structure of the protein, its folding and function.

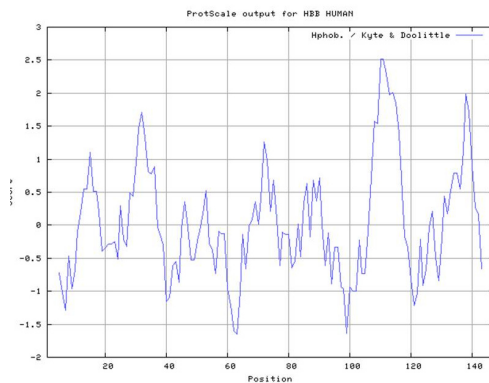
The tools in Uniprot are integrated in the database next to the sequence information. There is for example BLAST, that works similarly to the tools explained in the Ensembl section. ProtParam computes various physical and chemical properties for a given protein sequence (known protein or unknown protein sequence) [132]. The computed parameters include the molecular weight, theoretical pI (isoelectric point), amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY).

GRAVY measures hydrophobicity. Increasing positive score indicates greater hydrophobicity. The extinction coefficient is a measurement of how strongly a protein absorbs light at a given wavelength. The instability is a crude estimate of the protein stability in a test tube (index below 40, the protein is usually stable). The half-life is a prediction of the time it takes for half of the amount of the protein present in a cell to disappear completely after its synthesis.



Information obtained from protParam [132].

The ProtScale [132] represents the desired physicochemical property (hydrophobicity, bulkiness, polarity) along the protein sequence, based on the properties of the amino acids. It is useful to compare these graphs with the predicted or experimental structure of a protein. Finally, the Peptide Cutter predicts potential substrate cleavage sites, cleaved by proteases or chemicals in a given protein sequence. The tool returns the query sequence with the possible cleavage sites mapped on it and/or a table of cleavage site positions. Helps in the design of experiments and in the identification of unknown sequences. The output shows the number of peptides generated in a protein digestion, and the size of each fragment.



Name of enzyme	No. of cleavages	Positions of cleavage sites
Arg-C proteinase	9	71 77 88 99 105 109 112 139 145
Asp-N endopeptidase	4	27 54 85 110
CNBr	2	1 144
Formic acid	4	28 55 86 111
Glutamyl endopeptidase	10	33 43 50 65 67 84 127 134 138 146
Pepsin (pH1.3)	46	2 3 6 12 12 13 14 14 15 20 21 30 31 37 40 41 46 50 51 56 57 69 75 76 78 80 89 96 97 106 106 113 115 116 117 119 120 132 133 135 136 136 137 140 148
Proteinase K	70	3 4 7 8 11 12 13 14 15 16 18 21 22 29 30 31 34 35 37 38 41 42 46 47 49 51 52 53 54 56 57 58 59 63 64 68 69 72 73 74 75 76 79 80 81 90 92 93 94 97 103 104 106 107 110 114 115 117 118 120 125 128 132 133 136 137 140 142 143 148
Thermolysin	46	2 6 10 11 12 13 14 15 17 20 21 30 34 37 40 46 52 56 57 58 68 71 72 74 75 78 89 91 93 96 102 103 105 106 109 113 116 117 119 124 132 135 136 139 142 143
Trypsin	21	26 36 45 61 66 71 77 85 88 99 101 105 108 109 112 126 139 141 145 147 150

ProtScale (left) and PeptideCutter (right) tools integrated in Uniprot. Information obtained from Uniprot [131].

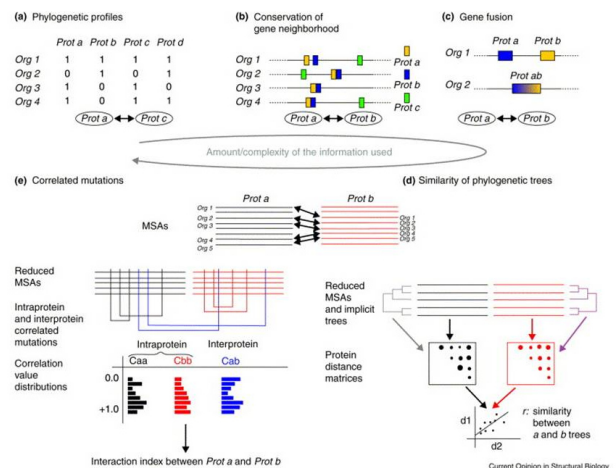
PROTEIN-PROTEIN INTERACTION DATABASES

There are two main approaches for detecting interacting proteins: techniques that measure direct physical interactions between protein pairs (for example Yeast to Hybrid-Y2H) and those that measure interactions among groups of proteins that may not form physical contacts – co-complex methods (usually involving protein coprecipitation and mass spectrometry).

In addition to these experimental methods, researchers have used computational techniques to predict interactions on the basis of factors such as amino-acid sequence and structural information.

Method	High-Throughput Approach	Living Cell Assay	Type of Interactions	Type of Characterization
Y2H [47,48]	+	In vivo	Physical interactions (binary)	Identification
Affinity purification-MS [61]	+	In vitro	Physical interactions (complex)	Identification
DNA microarrays/Gene coexpression [113]	+	In vitro	Functional association	Identification
Protein microarrays [114-116]	+	In vitro	Physical interaction (complex)	Identification
Synthetic lethality [85,86]	+	In vivo	Functional association	Identification
Phage display [117]	+	In vitro	Physical interaction (complex)	Identification
X-ray crystallography, NMR spectroscopy [84]	-	In vitro	Physical interactions (complex)	Structural and biological characterization
Fluorescence resonance energy transfer [89]	-	In vivo	Physical interaction (binary)	Biological characterization
Surface plasmon resonance [91]	-	In vitro	Physical interaction (complex)	Kinetic, dynamic characterization
Atomic force microscopy [93]	-	In vitro	Physical interaction (binary)	Mechanical, dynamic characterization
Electron microscopy [118]	-	In vitro	Physical interaction (complex)	Structural and biological characterization

High-throughput techniques are indicated with pluses (second column), and those which can provide information on interactions in vivo are shown in the third column. Fourth column indicates whether the method supplies data on physically interacting proteins in a complex ("complex") or only pairwise interactions ("binary"). Methods inferring interactions through functional association are shown as well. The type of protein interaction characterization is shown in the last column.
doi:10.1371/journal.ppat.0030042.t001



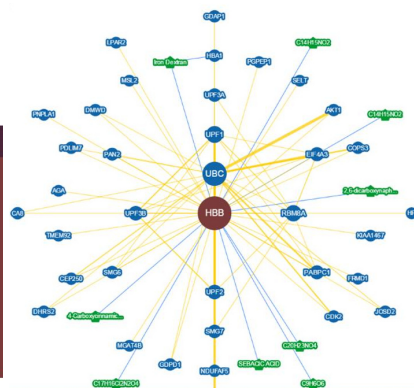
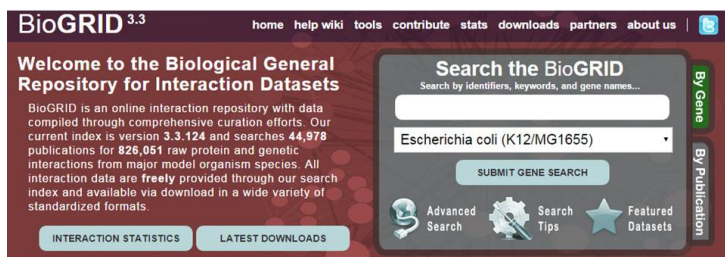
Left, Table showing techniques used to obtain protein protein data, from B. A. Shoemaker et al. [133]. Right, methods for predicting protein interaction partners from genomic and sequence information: (a) phylogenetic profiles, (b) Conservation of gene neighborhood (c) Gene fusion and (d) Similarity of phylogenetic trees. Figure from A. Valencia [134].

For example, in the Y2H method, the two proteins are tagged with half of a transcription factor that activates a desired gene. If the proteins interact physically, the transcription factor activates the gene, and we can measure the output. The specificity of this method is low, giving many false positives. The solutions are combining results with other methods or with data available in public databases (for example using known PPIs networks of other organisms, co-expression data, and bioinformatics tools for identification of sequences in the proteins that promote specific interactions between proteins).

THE BIOGRID

BioGRID is an interaction repository with data compiled through comprehensive curation efforts [135]. BioGRID interactions are recorded as relationships between two proteins or genes (i.e., they are binary relationships) with an evidence code that supports the interaction and a publication reference. The term "interaction" includes both physical binding of two proteins as well as co-existence in a stable complex and genetic interaction. It should not be assumed that the interaction reported in BioGRID is direct and physical in nature. Some interactions in BioGRID have various levels of evidential support. Therefore, the interactors are the proteins/molecules that interact, while the interactions refers to the number of evidences supporting the relationship between two interactors. The BioGRID Network viewer is a graphical tool for visualizing complex networks. You can access it from any of our result pages by simply choosing the "network" view.

<http://thebiogrid.org/>



The Biogrid panel and interactions from HBB. Screenshots from <http://thebiogrid.org> [136].

The information that can be found in the Biogrid is based on different fields. First, standard annotation data for search results (official name, synonyms, organism name, and description), as well as the functional information for the search result and external links to other databases. Second, a quick summary of the number and type (physical or genetic) of interactions represented in the current results. The colored bar behind each number denotes the relationship between high throughput and low throughput interactions. Third, proteins and molecules (interactors) for your search result, including drugs (green nodes) and proteins from other organisms (yellow nodes). Annotation data for each of the resulting is also provided. And forth, number and links to the supporting evidence from the different associations. There are links to the publications supporting these associations.

STRING-DB. DETERMINING THE FUNCTION BY COMPLEMENTATION

A different interaction database in STRING [137]. STRING has been developed by the Swiss Institute of Bioinformatics, together with the EMBL and the CPR (Centre for Protein Research). It is a database for known and predicted protein-protein interactions, including direct (physical) and indirect (functional). It has information on interactions from many different organisms (more than 2000), thanks to search from numerous sources including experimental data, computational prediction methods and public text collections.

In STRING we can generate interaction networks of a single protein, or a list of multiple proteins (for example, the interaction network of the top 20 mutated genes in cancer is shown below). It can also search interactions with the amino acid sequence input of a protein. The network view summarizes the network of predicted associations for a particular group of proteins. The network nodes are proteins. There are two modes in STRING. In the evidence mode, the color of the edges represent the predicted functional associations (experimental evidence, database evidence, textmining evidence). In the confidence mode, the thickness of the edge represents the confidence (score) of the predicted interaction.

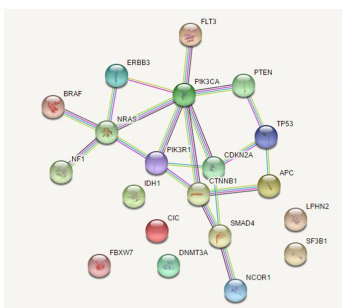


Image of STRING generated network [128].

The analysis tab allows to perform an enrichment of the biological processes or molecular functions or pathways of the network. Therefore, you can predict by complementation the function of an unknown protein by viewing the analysis tab and knowing the processes involved in the interacting proteins.

METABOLITE DATABASES

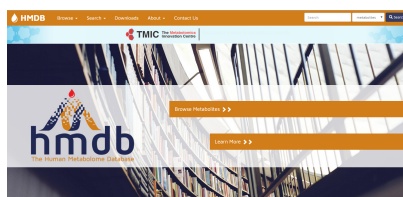
Metabolites are the intermediates and products of metabolism. Examples include antibiotics, pigments, carbohydrates, fatty acids and amino acids. There are primary (essential for living) and secondary metabolites. There is a wide variety of metabolite and metabolite levels in the human organism (and in different organisms). This makes the database approach the most challenging one, due to the great number of different metabolites and the diversity in properties, wide range of concentrations, analytical methods to detect them. Therefore, there are many very specific databases for different types of metabolites (lipids – LipidMaps-, glycans –UniCarbKB-), types of experiments (MS or NMR), or organisms (humans, bacteria, yeast, plants...). An example of different metabolite databases and repositories can be seen below.

Database name	URL or web address	Comments
Human metabolome database	http://www.hmdb.ca	Largest and most complete of its kind. Specific to humans only
BioMagResBank (BMRB – metabolomics)	http://www.bmrwisc.edu/metabolomics/	Emphasis on NMR data, no biological or biochemical data
BiGG (database of biochemical, genetic and genomic metabolic network reconstructions)	http://bigg.ucsd.edu/home.pl	Specific to plants (Arabidopsis) Database of human, yeast and bacterial metabolites, pathways and reactions as well as SBML reconstructions for metabolic modeling
Fiehn metabolome database	http://fiehnlab.ucdavis.edu/compounds/	Tabular list of ID ¹ metabolites with images, synonyms and KEGG links
Golm metabolome database	http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html	Emphasis on MS or GC-MS data only No biological data Few data fields Specific to plants
METLIN metabolite database	http://metlin.scripps.edu/	Human specific Mixes drugs, drug metabolites together Name, structure, ID only
NIST spectral database	http://webbook.nist.gov/chemistry/	Spectral database only (NMR, MS, IR) No biological data, little chemical data Not limited to metabolites
Spectral database for organic compounds (SDBS)	http://www.aist.go.jp/RIODB/SDBS/cgi-bin/direct.frameset.cgi?lang=eng	Spectral database only (NMR, MS, IR) No biological data, little chemical data Not limited to metabolites

Summary of metabolomic databases.

HUMAN METABOLOME DATABASE (HMDB)

As an example of metabolite database, and due to its importance to biomedical engineers, we are going to focus on the the Human Metabolome Database (HMDB) is a freely available electronic database containing detailed information about small molecule metabolites found in the human body [138]. It is part of the Human Metabolome Project (www.hmdb.ca). It is intended to be used for applications in metabolomics, clinical chemistry, biomarker discovery and general education. The database (version 3.6) contains 68,265 metabolite entries including both water-soluble and lipid soluble metabolites as well as metabolites that would be regarded as either abundant (> 1 μ M) or relatively rare (< 1 nM).



HMDB webpage. Screenshot from <https://hmdb.ca>

The database is designed to contain or link three kinds of data: 1) chemical data, 2) clinical data, and 3) molecular biology/biochemistry data. In general, the HMDB provides, for each metabolite, information about metabolite physicochemical properties, its function, the structure (3D downloadable PyMOL structure), synonyms, formula, molecular weight. It also contains links to metabolic pathways and related enzymes, a taxonomical classification of the metabolite, its localization in the organism and its biological properties.

Additionally, it has links to experiments where the metabolite concentration in both normal and abnormal conditions have been measured in different biofluids (blood, urine, saliva, breast milk, cerebrospinal fluid...). Therefore, it is an easily accessible tool to find information about possible metabolite biomarkers of disease.

REFERENCES

- [1] H. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nature Structural Biology*, vol. 10, no. 12, p. 980, 2003.
- [2] M. Y. Galperin, X. M. Fernández-Suárez, and D. J. Rigden, "The 24th annual Nucleic Acids Research database issue: A look back and upcoming changes," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D1–D11, 2017.
- [3] S. Dunin-Horkawicz, K. O. Kopec, and A. N. Lupas, "Prokaryotic ancestry of eukaryotic protein networks mediating innate immunity and apoptosis," *J. Mol. Biol.*, vol. 426, no. 7, pp. 1568–1582, 2014.
- [4] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nat Rev Genet*, vol. 17, no. 6, pp. 333–351, 2016.
- [5] R. P. Hertzberg and A. J. Pope, "High-throughput screening: New technology for the 21st century," *Current Opinion in Chemical Biology*, vol. 4, no. 4, pp. 445–451, 2000.
- [6] J. North, *Cosmos: an illustrated history of astronomy and cosmology*. University of Chicago Press, 2008.
- [7] A. Rad, "Wikimedia commons," *Wikimedia commons*. 2007.
- [8] P. J. Kraulis, "MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures," *J. Appl. Crystallogr.*, vol. 24, no. 5, pp. 946–950, 1991.
- [9] C. E. Ophardt, "Virtual chembook," *Elmhurst Coll.*, pp. 121–125, 2003.
- [10] R. J. Gillespie, "The valence-shell electron-pair repulsion (VSEPR) theory of directed valency," *J. Chem. Educ.*, vol. 40, no. 6, p. 295, 1963.
- [11] C. C. J. Roothaan, "New developments in molecular orbital theory," *Rev. Mod. Phys.*, 1951.
- [12] M. R. Leach, "www.chemistry-drills.com." [Online]. Available: <http://www.chemistry-drills.com/functional-groups.php?q=simple>. [Accessed: 19-Sep-2009].
- [13] James, "MasterOrganicChemistry.com." [Online]. Available: <https://www.masterorganicchemistry.com/2010/10/06/functional-groups-organic-chemistry/>.
- [14] R. Guha *et al.*, "The Blue Obelisk – interoperability in chemical informatics," *J. Chem. Inf. Model.*, vol. 46, no. 3, pp. 991–998, 2006.
- [15] N. C. for Biotechnology, "PubChem Substructure Fingerprint V1.3," 2009. [Online]. Available: <http://pubchem.ncbi.nlm.nih.gov>.
- [16] T. Sterling and J. J. Irwin, "ZINC15–Ligand Discovery for Everyone," *J. Chem. Inf. Model.*, vol. 55, no. 11, pp. 2324–2337, 2015.
- [17] A. R. Leach, "Molecular modelling : principles and applications," *Computers*, vol. 21, no. 3, p. 784, 2001.
- [18] D. V Svintradze, "Moving Manifolds in Electromagnetic Fields," *Front. Phys.*, vol. 5, p. 37, 2017.
- [19] J. Klett *et al.*, "MM-ISMSA: An ultrafast and accurate scoring function for protein-protein docking," *J. Chem. Theory Comput.*, vol. 8, no. 9, pp. 3395–3408, 2012.
- [20] K. P. Tan, T. B. Nguyen, S. Patel, R. Varadarajan, and M. S. Madhusudhan, "Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins," *Nucleic Acids Res.*, 2013.
- [21] S. Genheden and U. Ryde, "The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities," *Expert Opin. Drug Discov.*, vol. 10, no. 5, pp. 449–461, 2015.
- [22] K. K. Frederick, M. S. Marlow, K. G. Valentine, and A. J. Wand, "Conformational entropy in molecular recognition by proteins," *Nature*, vol. 448, no. 7151, pp. 325–329, 2007.
- [23] A. J. Wand and K. A. Sharp, "Measuring Entropy in Molecular Recognition by Proteins.," *Annu. Rev. Biophys.*, 2018.

- [24] H. B. Schlegel, "Optimization of equilibrium geometries and transition structures," *J. Comput. Chem.*, vol. 3, no. 2, pp. 214–218, 1982.
- [25] E. F. Pettersen *et al.*, "UCSF Chimera – A Visualization System for Exploratory Research and Analysis," *J. Comput. Chem.*, vol. 25, pp. 1605–1612, 2004.
- [26] G. Maggiora, M. Vogt, D. Stumpfe, and J. Bajorath, "Molecular similarity in medicinal chemistry," *Journal of Medicinal Chemistry*, vol. 57, no. 8, pp. 3186–3204, 2014.
- [27] A. H. Lipkus, "A proof of the triangle inequality for the Tanimoto distance," *J. Math. Chem.*, vol. 26, no. 1–3, pp. 263–265, 1999.
- [28] M. S. Armstrong *et al.*, "ElectroShape: Fast molecular similarity calculations incorporating shape, chirality and electrostatics," *J. Comput. Aided. Mol. Des.*, vol. 24, no. 9, pp. 789–801, 2010.
- [29] J. Klett, Á. Cortés-Cabrera, R. Gil-Redondo, F. Gago, and A. Morreale, "ALFA: Automatic ligand flexibility assignment," *J. Chem. Inf. Model.*, vol. 54, no. 1, pp. 314–323, 2014.
- [30] M. J. Hartshorn *et al.*, "Diverse, high-quality test set for the validation of protein-ligand docking performance," *J. Med. Chem.*, vol. 50, no. 4, pp. 726–741, 2007.
- [31] S.-Y. Yang, "Pharmacophore modelling and applications in drug discovery: challenges and recent advances," *Drug Discov. Today*, vol. 15, no. 11–12, pp. 444–450, 2010.
- [32] A. V. Grigoryan, I. Kufareva, M. Totrov, and R. A. Abagyan, "Spatial chemical distance based on atomic property fields," *J. Comput. Aided. Mol. Des.*, vol. 24, no. 3, pp. 173–182, 2010.
- [33] A. J. Williams, "Public Chemical Compound Databases," *Curr. Opin. Drug Discov. Devel.*, vol. 11, no. 3, pp. 393–404, 2008.
- [34] D. S. Wishart *et al.*, "DrugBank 5.0: A major update to the DrugBank database for 2018," *Nucleic Acids Res.*, 2018.
- [35] A. Gaulton *et al.*, "The ChEMBL database in 2017," *Nucleic Acids Res.*, 2017.
- [36] M. Awale, R. Van Deursen, and J. L. Reymond, "MQN-mapplet: Visualization of chemical space with interactive maps of DrugBank, ChEMBL, PubChem, GDB-11, and GDB-13," *J. Chem. Inf. Model.*, vol. 53, no. 2, pp. 509–518, 2013.
- [37] R. A. Dwek, "Glycobiology: Toward Understanding the Function of Sugars," *Chem. Rev.*, vol. 96, no. 2, pp. 683–720, 1996.
- [38] P. M. Rudd and R. A. Dwek, "Glycosylation: Heterogeneity and the 3D structure of proteins," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 32, no. 1, pp. 1–100, 1997.
- [39] K. N. Kirschner *et al.*, "GLYCAM06: A generalizable biomolecular force field. carbohydrates," *J. Comput. Chem.*, vol. 29, no. 4, pp. 622–655, 2008.
- [40] C. Nishida and N. Fenandez Martinez, "The Role of Carbohydrates in Nutrition: Joint FAO/WHO Expert Consultation," *Food Nutr. Pap.*, vol. 66, pp. 1–129, 1998.
- [41] K. Nagy and I.-D. Tiuca, "Importance of Fatty Acids in Physiopathology of Human Body," in *Fatty Acids*, 2017.
- [42] S. Jo, J. B. Lim, J. B. Klauda, and W. Im, "CHARMM-GUI membrane builder for mixed bilayers and its application to yeast membranes," *Biophys. J.*, vol. 97, no. 1, pp. 50–58, 2009.
- [43] D. P. Tieleman and H. J. C. Berendsen, "Molecular dynamics simulations of a fully hydrated dipalmitoylphosphatidylcholine bilayer with different macroscopic boundary conditions and parameters," *J. Chem. Phys.*, vol. 105, no. 11, pp. 4871–4880, 1996.
- [44] D. P. Tieleman, D. van der Spoel, and H. J. C. Berendsen, "Molecular Dynamics Simulations of Dodecylphosphocholine Micelles at Three Different Aggregate Sizes: Micellar Structure and Chain Relaxation," *J. Phys. Chem. B*, vol. 104, no. 27, pp. 6380–6388, 2000.
- [45] D. P. Tieleman, M. S. P. Sansom, and H. J. C. Berendsen, "Alamethicin Helices in a Bilayer and in Solution: Molecular Dynamics Simulations," *Biophys. J.*, vol. 76, no. 1, pp. 40–49, 1999.

- [46] A. T. Annunziato, "DNA Packaging: Nucleosomes and Chromatin," *Nat. Educ.*, vol. 1, no. 1, p. 26, 2008.
- [47] B. Wu, K. Mohideen, D. Vasudevan, and C. A. Davey, "Structural Insight into the Sequence Dependence of Nucleosome Positioning," *Structure*, 2010.
- [48] W. R. Pearson, "Rapid and sensitive sequence comparison with FASTP and FASTA," *Methods Enzymol.*, vol. 183, no. C, pp. 63–98, 1990.
- [49] S. Kumar, G. Stecher, and K. Tamura, "MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets," *Mol. Biol. Evol.*, vol. 33, no. 7, pp. 1870–1874, 2016.
- [50] P. E. Wright and H. J. Dyson, "Intrinsically disordered proteins in cellular signalling and regulation," *Nat. Rev. Mol. Cell Biol.*, vol. 16, no. 1, pp. 18–29, 2014.
- [51] J. J. Ward, L. J. McGuffin, K. Bryson, B. F. Buxton, and D. T. Jones, "The DISOPRED server for the prediction of protein disorder," *Bioinformatics*, vol. 20, no. 13, pp. 2138–2139, 2004.
- [52] J. A. Marsh and S. A. Teichmann, "Structure, Dynamics, Assembly, and Evolution of Protein Complexes," *Annu. Rev. Biochem.*, vol. 84, no. 1, pp. 551–575, 2015.
- [53] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, "Stereochemistry of polypeptide chain configurations," *Journal of Molecular Biology*, vol. 7, no. 1, pp. 95–99, 1963.
- [54] R. W. Hooft, C. Sander, and G. Vriend, "Objectively judging the quality of a protein structure from a Ramachandran plot," *Comput. Appl. Biosci.*, vol. 13, no. 4, pp. 425–430, 1997.
- [55] W. L. DeLano, "The PyMOL Molecular Graphics System, Version 2.3," *Schrödinger LLC*. 2020.
- [56] M. S. Smyth and J. H. J. Martin, "x Ray crystallography," *Journal of Clinical Pathology - Molecular Pathology*. 2000.
- [57] Martín Martínez-Ripoll, "Crystallography-Cristalografia." [Online]. Available: <https://www.xtal.iqfr.csic.es/Cristalografia/index-en.html>.
- [58] N. Karpukhina, R. G. Hill, and R. V. Law, "Crystallisation in oxide glasses-a tutorial review," *Chemical Society Reviews*. 2014.
- [59] S. K. Burley *et al.*, "RCSB Protein Data Bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy," *Nucleic Acids Res.*, 2019.
- [60] G. J. Kleywegt and T. A. Jones, "Model building and refinement practice," *Methods Enzymol.*, 1997.
- [61] K. Murata and M. Wolf, "Cryo-electron microscopy for structural analysis of dynamic biological macromolecules," *Biochimica et Biophysica Acta - General Subjects*. 2018.
- [62] M. Parmar *et al.*, "Using a SMALP platform to determine a sub-nm single particle cryo-EM membrane protein structure," *Biochim. Biophys. Acta - Biomembr.*, 2018.
- [63] E. Callaway, "The revolution will not be crystallized: A new method sweeps through structural biology," *Nature*, 2015.
- [64] W. Kühlbrandt, "The resolution revolution," *Science*. 2014.
- [65] A. R. Faruqi and R. Henderson, "Electronic detectors for electron microscopy," *Current Opinion in Structural Biology*. 2007.
- [66] A. Pryor *et al.*, "GENFIRE: A generalized Fourier iterative reconstruction algorithm for high-resolution 3D imaging," *arXiv*. 2017.
- [67] A. Loquet, J. Tolchard, M. Berbon, D. Martinez, and B. Habenstein, "Atomic scale structural studies of macromolecular assemblies by solid-state nuclear magnetic resonance spectroscopy," *J. Vis. Exp.*, vol. 2017, no. 127, pp. 1–12, 2017.
- [68] G. M. Clore and A. M. Gronenborn, "Determination of three-dimensional structures of proteins and nucleic acids in solution by nuclear magnetic resonance spectroscopy.," *Crit. Rev. Biochem. Mol. Biol.*, vol. 24, no. 5, pp. 479–564, 1989.
- [69] J. M. Berg, J. L. Tymoczko, and L. Stryer, "Biochemistry. 5th edition.," in *Biochemistry textbook*, 2006, p. 1120.

- [70] Y. Xu *et al.*, "Structural insight into SUMO chain recognition and manipulation by the ubiquitin ligase RNF4," *Nat. Commun.*, 2014.
- [71] C. Bissantz, B. Kuhn, and M. Stahl, "A medicinal chemist's guide to molecular interactions," *Journal of Medicinal Chemistry*. 2010.
- [72] A. Shrake and J. A. Rupley, "Environment and exposure to solvent of protein atoms. Lysozyme and insulin," *J. Mol. Biol.*, vol. 79, no. 2, 1973.
- [73] C. Zahnd, S. Spinelli, B. Luginbühl, P. Amstutz, C. Cambillau, and A. Plückthun, "Directed in Vitro Evolution and Crystallographic Analysis of a Peptide-binding Single Chain Antibody Fragment (scFv) with Low Picomolar Affinity," *J. Biol. Chem.*, 2004.
- [74] J. A. Reynolds, D. B. Gilbert, and C. Tanford, "Empirical Correlation Between Hydrophobic Free Energy and Aqueous Cavity Surface Area," *Proc. Natl. Acad. Sci.*, 1974.
- [75] G. B. McGaughey, M. Gagné, and A. K. Rappé, " π -Stacking interactions. Alive and well in proteins," *J. Biol. Chem.*, vol. 273, no. 25, pp. 15458-15463, 1998.
- [76] D. Kesters *et al.*, "Structural basis of ligand recognition in 5-HT₃ receptors," *EMBO Rep.*, 2013.
- [77] Y. Fukunishi, D. Mitomo, and H. Nakamura, "Protein-ligand binding free energy calculation by the Smooth Reaction Path Generation (SRPG) method," *J. Chem. Inf. Model.*, vol. 49, no. 8, pp. 1944-1951, 2009.
- [78] A. Cherkasov *et al.*, "QSAR modelling: Where have you been? Where are you going to?," *Journal of Medicinal Chemistry*. 2014.
- [79] P. Csermely, R. Palotai, and R. Nussinov, "Induced fit, conformational selection and independent dynamic segments: An extended view of binding events," *Trends Biochem. Sci.*, vol. 35, no. 10, pp. 539-546, 2010.
- [80] D. E. Koshland, "The Key-Lock Theory and the Induced Fit Theory," *Angewandte Chemie International Edition in English*, vol. 33, no. 23-24, pp. 2375-2378, 1995.
- [81] J. E. Straub and D. Thirumalai, "Theoretical probes of conformational fluctuations in S-peptide and RNase A/3'-UMP enzyme product complex," *Proteins Struct. Funct. Bioinforma.*, vol. 15, no. 4, pp. 360-373, 1993.
- [82] D. S. Goodsell, G. M. Morris, and A. J. Olson, "Automated docking of flexible ligands: Applications of AutoDock," *J. Mol. Recognit.*, 1996.
- [83] C. H. Weber and C. Vincenz, "A docking model of key components of the DISC complex: Death domain superfamily interactions redefined," *FEBS Lett.*, 2001.
- [84] D. Xu, K. Baburaj, C. B. Peterson, and Y. Xu, "Model for the three-dimensional structure of vitronectin: Predictions for the multi-domain protein from threading and docking," *Proteins Struct. Funct. Genet.*, 2001.
- [85] D. Kozakov *et al.*, "The ClusPro web server for protein-protein docking," *Nat. Protoc.*, 2017.
- [86] B. Jiménez-García, C. Pons, and J. Fernández-Recio, "pyDockWEB: A web server for rigid-body protein-protein docking using electrostatics and desolvation scoring," in *Bioinformatics*, 2013.
- [87] E. C. Meng, B. K. Shoichet, and I. D. Kuntz, "Automated docking with grid-based energy evaluation," *J. Comput. Chem.*, 1992.
- [88] D. Whitley, "A genetic algorithm tutorial," *Stat. Comput.*, vol. 4, no. 2, pp. 65-85, 1994.
- [89] Morris G.M. and Dallakyan S., "AutoDock – AutoDock," 02-27, vol. 1, no. 1, pp. 15-45, 2013.
- [90] R. A. Friesner *et al.*, "Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes," *J. Med. Chem.*, 2006.
- [91] P. Schmidtke, V. Le Guilloux, J. Maupetit, and P. Tufféry, "fpocket: Online tools for protein ensemble pocket detection and tracking," *Nucleic Acids Res.*, 2010.
- [92] A. J. Trott, O., Olson, "Autodock vina: improving the speed and accuracy of docking," *J. Comput. Chem.*, 2010.
- [93] C. Regnault, D. S. Dheeman, and A. Hochstetter, "Microfluidic devices for drug assays," *High-Throughput*. 2018.

- [94] G. Harper and S. D. Pickett, "Methods for mining HTS data," *Drug Discovery Today*, vol. 11, no. 15–16, pp. 694–699, 2006.
- [95] M. Urbano-Cuadrado, O. Rabal, and J. Oyarzabal, "Centralizing Discovery Information: From Logistics to Knowledge at a Public Organization," *Comb. Chem. High Throughput Screen.*, 2011.
- [96] L. Zemanová, A. Schenk, M. J. Valler, G. U. Nienhaus, and R. Heilker, "Confocal optics microscopy for biochemical and cellular high-throughput screening," *Drug Discovery Today*. 2003.
- [97] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.*, vol. 48, no. 3, pp. 443–453, 1970.
- [98] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol.*, 1981.
- [99] V. Modi and R. L. Dunbrack, "Defining a new nomenclature for the structures of active and inactive kinases," *Proc. Natl. Acad. Sci. U. S. A.*, 2019.
- [100] C. Chothia and A. M. Lesk, "The relation between the divergence of sequence and structure in proteins," *EMBO J.*, vol. 5, no. 4, pp. 823–6, 1986.
- [101] Á. Sebastián Yagüe *et al.*, "' Bioinformática con Ñ v1. 0': a collaborative project of young Spanish scientists to write a complete book about Bioinformatics," 2014.
- [102] D. W. Mount, "Comparison of the PAM and BLOSUM amino acid substitution matrices," *Cold Spring Harb. Protoc.*, 2008.
- [103] S. F. Altschul *et al.*, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.
- [104] S. A. Bukhari and G. Caetano-Anollés, "Origin and Evolution of Protein Fold Designs Inferred from Phylogenomic Analysis of CATH Domain Structures in Proteomes," *PLoS Comput. Biol.*, 2013.
- [105] C. Pandya *et al.*, "Consequences of domain insertion on sequence-structure divergence in a superfold," *Proc. Natl. Acad. Sci. U. S. A.*, 2013.
- [106] Y. Haddad, V. Adam, and Z. Heger, "Ten quick tips for homology modelling of high-resolution protein 3D structures," *PLoS Computational Biology*. 2020.
- [107] A. Cembran, J. Kim, J. Gao, and G. Veglia, "NMR mapping of protein conformational landscapes using coordinated behavior of chemical shifts upon ligand binding," *Phys. Chem. Chem. Phys.*, 2014.
- [108] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, "Anisotropy of fluctuation dynamics of proteins with an elastic network model," *Biophys. J.*, vol. 80, no. 1, pp. 505–515, 2001.
- [109] P. Campitelli and S. B. Ozkan, "Allostery and epistasis: Emergent properties of anisotropic networks," *Entropy*, 2020.
- [110] M. M. Tirion, "Large amplitude elastic motions in proteins from a single-parameter, atomic analysis," *Phys. Rev. Lett.*, 1996.
- [111] G. Hu, L. He, F. Iacovelli, and M. Falconi, "Intrinsic Dynamics Analysis of a DNA Octahedron by Elastic Network Model," *Molecules*, vol. 22, no. 1, 2017.
- [112] W. G. Krebs, V. Alexandrov, C. A. Wilson, N. Echols, H. Yu, and M. Gerstein, "Normal mode analysis of macromolecular motions in a database framework: Developing mode concentration as a useful classifying statistic," *Proteins Struct. Funct. Genet.*, 2002.
- [113] B. Surpeta, C. E. Sequeiros-Borja, and J. Brezovsky, "Dynamics, a powerful component of current and future in silico approaches for protein design and engineering," *International Journal of Molecular Sciences*. 2020.
- [114] M. Goguet, T. J. Narwani, R. Petermann, V. Jallu, and A. G. De Brevern, "In silico analysis of Glanzmann variants of Calf-1 domain of $\alpha\text{IIb}\beta\text{3}$ integrin revealed dynamic allosteric effect," *Sci. Rep.*, 2017.
- [115] C. Alcocer-Cuarón, A. L. Rivera, and V. M. Castaño, "Hierarchical structure of biological systems: A bioengineering approach," *Bioengineered*. 2014.
- [116] D. J. Rigden and X. M. Fernández, "The 27th annual Nucleic Acids Research database issue and molecular biology database collection," *Nucleic Acids Res.*, 2020.

- [117] M. Chagoyen Quiles, "Integration of biological data: systems, infrastructures and programmable tools," 2005.
- [118] S. Burge *et al.*, "Biocurators and biocuration: Surveying the 21st century challenges," *Database*, 2012.
- [119] A. D. Baxevanis, "The importance of biological databases in biological discovery," *Curr. Protoc. Bioinforma.*, 2011.
- [120] R. Garcia-Milian, D. Hersey, M. Vukmirovic, and F. Duprilot, "Data challenges of biomedical researchers in the age of omics," *PeerJ*, 2018.
- [121] Z. D. Stephens *et al.*, "Big data: Astronomical or genomics?," *PLoS Biol.*, 2015.
- [122] Q. Chen, J. Zobel, and K. Verspoor, "Duplicates, redundancies and inconsistencies in the primary nucleotide databases: A descriptive study," *Database*, 2017.
- [123] S. E. Hunt *et al.*, "Ensembl variation resources," *Database (Oxford)*, 2018.
- [124] J. S. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh, "OMIM.org: Leveraging knowledge across phenotype-gene relationships," *Nucleic Acids Res.*, 2019.
- [125] S. T. Sherry *et al.*, "dbSNP: The NCBI database of genetic variation," *Nucleic Acids Res.*, 2001.
- [126] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, 1990.
- [127] J. Chang, M. W. Carrillo, A. Waugh, L. Wei, and R. B. Altman, "Scoring Functions Sensitive to Alignment Error Have a More Difficult Search: A Paradox for Threading," *ACS Symp. Ser.*, 2002.
- [128] The 1000 Genomes Project, "ARTICLE A global reference for human genetic variation The 1000 Genomes Project Consortium*," *Nat. Artic.*, 2015.
- [129] W. McLaren *et al.*, "The Ensembl Variant Effect Predictor," *Genome Biol.*, 2016.
- [130] I. A. Adzhubei *et al.*, "A method and server for predicting damaging missense mutations," *Nature Methods*. 2010.
- [131] A. Bateman, "UniProt: A worldwide hub of protein knowledge," *Nucleic Acids Res.*, 2019.
- [132] E. Gasteiger *et al.*, "Protein Identification and Analysis Tools on the ExPASy Server," in *The Proteomics Protocols Handbook*, 2005.
- [133] B. A. Shoemaker and A. R. Panchenko, "Deciphering protein-protein interactions. Part I. Experimental techniques and databases," *PLoS Computational Biology*. 2007.
- [134] A. Valencia and F. Pazos, "Computational methods for the prediction of protein interactions," *Current Opinion in Structural Biology*. 2002.
- [135] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Res.*, 2006.
- [136] R. Oughtred *et al.*, "The BioGRID interaction database: 2019 update," *Nucleic Acids Res.*, 2019.
- [137] D. Szklarczyk *et al.*, "The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible," *Nucleic Acids Res.*, 2017.
- [138] D. S. Wishart *et al.*, "HMDB 4.0: The human metabolome database for 2018," *Nucleic Acids Res.*, 2018.