

Protein Folding

Archivos de Secuencias, de Estructuras, y PDB

Luis Garreta

Electiva de Bioinformática
MAESTRÍA EN INFORMÁTICA BIOMÉDICA
Universidad del Bosque
Bogotá-Colombia

February 26, 2021

Amino acid sequences codes

- There are 20 types of amino acids (AA) in proteins
- 1-letter code used in the AA sequence files

Amino Acid	3-Letters	1-Letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Amino acid sequence file:

Fasta files

Fasta format:

- The header in the first line
- Followed by the sequence

```
>sp|P26239|BCHI_RHOCB Magnesium-chelatase 38 kDa subunit OS=Rhodobacter capsulatus
MTTAVARLQPSASGAKTRPVFPFSAIVGQEDMKLALLLTAVDPGIGGVLVFGDRGTGKST
AVRALAALLPEIEAVEGCPVSSPNVEMIPDWATVLSTNVIRKPTPVVDLPLGVSEDRVVG
ALDIERAISKGEKAFEPGLLARANRGYLYIDECNLLEDHIVDLLDVAQSGENVVERDGL
SIRHPARFVLVGSGNPEEGDLRPQLLDRFGLSVEVLSPRDVETRVVIRRRDTYDADPKA
FLEEWRPKMDIRNQILEARERLPKVEAPNTALYDCAALCIALGSDGLRGELTLLRSARA
LAALEGATAVGRDHLKRVATMALSHRLRRDPLDEAGSTARVARTVEETLP
```

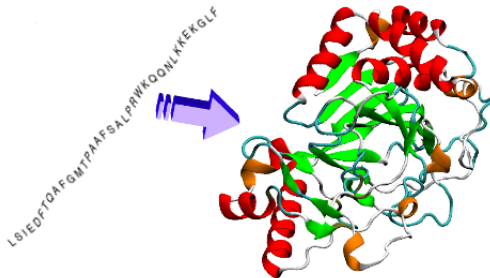
Multi Sequence Alignments

Comparative amino acid sequence alignment of POMT2 protein across various animal kingdoms shows the conservation of arginine amino acid during evolution (marked in the rectangle)



Proteins folds in 3D structures

Proteins fold from a linear sequence to a 3D structure

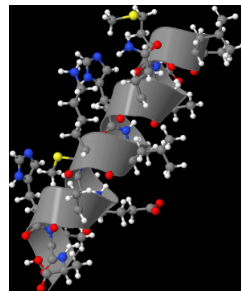


Protein Structures: PDB Files

Atomic Coordinates: PDB Format

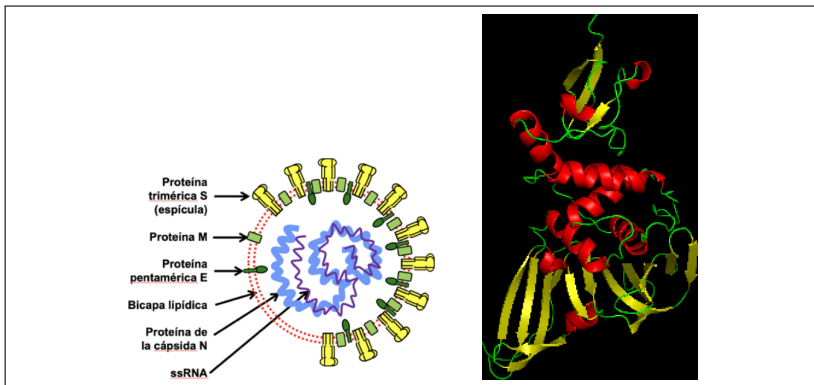
						Chain name			Sequence Number					
Amino Acid														
Element												-----Coordinates-----		
												X	Y	Z
ATOM	1	N	ASP	L	1	4.060	7.307	5.186						
ATOM	2	CA	ASP	L	1	4.042	7.776	6.553						
ATOM	3	C	ASP	L	1	2.668	8.426	6.644						
ATOM	4	O	ASP	L	1	1.987	8.438	5.606						
ATOM	5	CB	ASP	L	1	5.090	8.827	6.797						
ATOM	6	CG	ASP	L	1	6.338	8.761	5.929						
ATOM	7	OD1	ASP	L	1	6.576	9.758	5.241						
ATOM	8	OD2	ASP	L	1	7.065	7.759	5.948						

\\
Element position within amino acid



Folding Trajectories: MultiPDB Files

Molecular Dynamics Simulations Related to SARS-CoV-2



/home/lg/lgdocs/UNIBOSQUE/talks/02-Protein-Seq-3D-PDB/trajectory

Protein Data Bases

Type	Available database
DNA sequence database (primary)	EMBL ⁷ , GenBank ⁸ , DDBJ ⁹
Genomic database (primary)	GDB ¹⁰ , ACeDB, SGD
Protein sequence database (primary)	SWISS—PROT ¹¹
Protein structure database (primary)	PDB ¹² , CarbBank
Genomic database (secondary)	EcoWeb, RansFac
Protein sequence database (secondary)	TrEMBL ¹³ , PROSITE ¹⁴
Protein structure database (secondary)	DSSP, FSSP, HSSP
Literature database	DBcat, MEDLINE, SCI, HUMAT
Miscellaneous database	BIOCAT, GENDIAG, LIMB, TaxonomyTAED
Comprehensive database	BioWarehouse ¹⁵

Protein Data Bank (PDB)

- Single worldwide database and hundreds of secondary databases categorize the data differently.
- **Key resource** in the area of structural biology, stores 3D structural data of large biological molecules such as Proteins and Nucleic acids.
- Data is **submitted by** Biologists and Biochemists from all around the world to be freely accessible on internet via its member organizations' websites and is updated weekly.
- The **mission** is to maintain a single Protein Data Bank Archive of Macromolecular Structural data.

<https://www.rcsb.org/>