# 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

Protein Folding

## Amino acid sequences codes

- Thera 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	С	
Glutamic acid	Glu	Е	
Glutamine	Gln	Q G	
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 AVRALAALLPETEAVEGCPVSSPMVEMTPDWATVLSTNVIRKPTPVVDLPLGVSEDRVVG ALDIERAISKGEKAFEPGLLARANRGYLYIDECNLLEDHIVDLLLDVAQSGENVVERDGL SIRHPARFVLVGSGNPEEGDLRPQLLDRFGLSVEVLSPRDVETRVEVIRRRDTYDADPKA FLEEWRPKDMDIRNQILEARERLPKVEAPNTALYDCAALCIALGSDGLRGELTLLRSARA LAALEGATAVGRDHLKRVATMALSHRLRRDPLDEAGSTARVARTVEETLP

# Multi Sequence Alignments

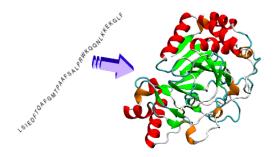
Comparative amino acid sequence alignment of POMT2 protein across various animal kingdoms shows the conservation of arginine amino acid during evolution (marked in the rectangle)

FruitFlv NHKTGGGYLHSHHHLYPKGSGARQQQVTTYTHKDENNKWLIRPHNKPG---PPKGKVQIL SLRTAAGYLHSHWHLYPEGVGARDOOVTTYSHKDONNLWIIKKHNSESGERDPASPVELV Xenopus Bovine NLRMAIGYLHSHRHFYPEGVGARDOOVTTYLHKDYNNLWIIKKHNSNSDPLDPSLPVEFV Mouse NLRMAIGYLHSHRHLYPEGIGARQQQVTTYLHKDYNNLWIIKKYNANTDPLDPSFPVEFV Elephant NLRMASGYLHSHRHLYPEGIGARQQQVTTYLHKDYNNLWIVKKHNTNSDPLDPSLPVEFV NLRMAIGYLHSHRHLYPEGIGARDOOVTTYLHKDYNNLWIIKKHNTNADPLDPSIPVEFV Dog Rabbit NLKMAIGYLHSHPHLYPEGVGARDOOVTTYLHKDYNNLWIIKKHNTDSDPLDPSSPVEFV NERMAIGYLHSHRHLYPEGIGARQQQVTTYLHKDYNNLWIIKKHNTNSDPLDPSFPVEFV Macaque NLRMAIGYLHSHRHLYPEGIGARQQQVTTYLHKDYNNLWIIKKHNTNSDPLDPSFPVEFV Chimpanzee NLRMAIGYLHSHRHLYPEGIGARQQQVTTYLHKDYNNLWIIKKHNTNSDPLDPSFPVEFV Human 

4 D > 4 A > 4 E > 4 E > E 904 A

### Proteins folds in 3D structures

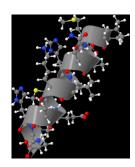
Proteins fold from a linear sequence to a 3D structure



## Protein Structures: PDB Files

#### **Atomic Coordinates: PDB Format**

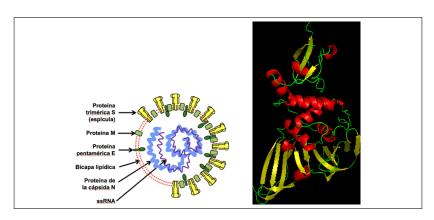
o Ac	id			/ 0-			
	١.		/ Sequence Number				
	•			/ /			
Element \			/	' /	Coordinates		
	\ '	\	1	/	x	Y	Z
1	N	ASP	L	1	4.060	7.307	5.186
2	CA	ASP	L	1	4.042	7.776	6.553
3	C	ASP	L	1	2.668	8.426	6.644
4	0	ASP	L	1	1.987	8.438	5.606
5	CB	ASP	L	1	5.090	8.827	6.797
6	CG	ASP	L	1	6.338	8.761	5.929
7	OD1	ASP	L	1	6.576	9.758	5.241
8	OD2	ASP	L	1	7.065	7.759	5.948
	1 2 3 4 5 6	1 N 2 CA 3 C 4 O 5 CB 6 CG 7 OD1	1 N ASP 2 CA ASP 3 C ASP 4 O ASP 5 CB ASP 6 CG ASP 7 OD1 ASP	1 N ASP L 2 CA ASP L 3 C ASP L 4 O ASP L 5 CB ASP L 6 CG ASP L 7 OD1 ASP L	1 N ASP L 1 2 CA ASP L 1 3 C ASP L 1 4 O ASP L 1 5 CB ASP L 1 6 CG ASP L 1 7 OD1 ASP L 1	1 N ASP L 1 4.060 2 CA ASP L 1 4.042 3 C ASP L 1 2.668 4 O ASP L 1 1.987 5 CB ASP L 1 5.090 6 CG ASP L 1 6.338 7 OD1 ASP L 1 6.576	1 N ASP L 1 4.060 7.307 2 CA ASP L 1 4.042 7.776 3 C ASP L 1 2.668 8.426 4 O ASP L 1 1.987 8.438 5 CB ASP L 1 5.090 8.827 6 CG ASP L 1 6.338 8.761 7 OD1 ASP L 1 6.576 9.758



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

# 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/

◆ロト→部ト→重ト→重・釣りで