

# Protein Sequence and Structure

Proteomes Interactomes and Biological Networks

Emidio Capriotti

<http://biofold.org/>

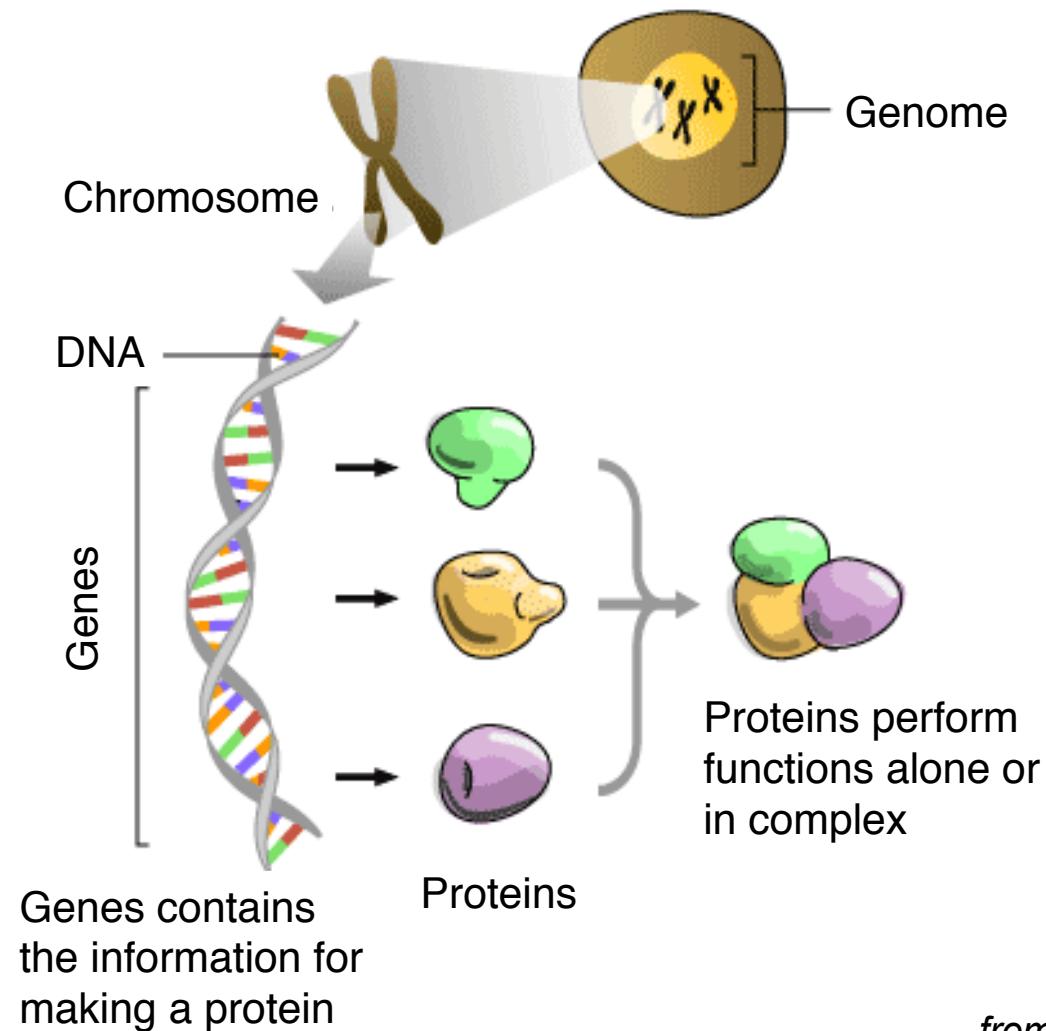


Biomolecules  
Folding and  
Disease

Department of Pharmacy and  
Biotechnology (FaBiT)  
University of Bologna



# The Central Dogma

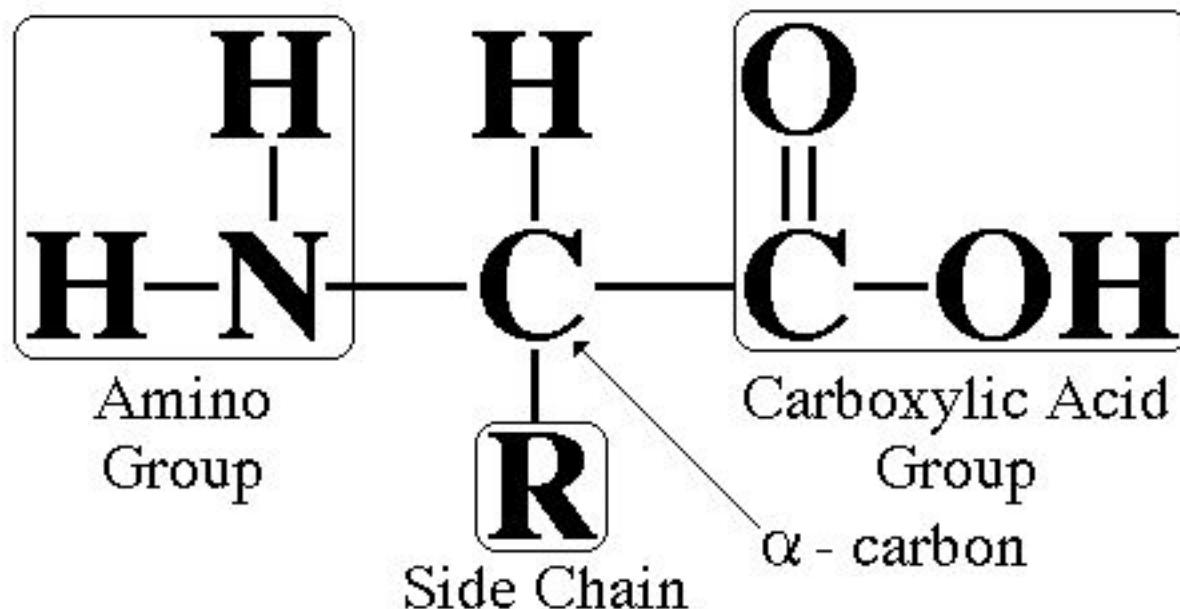


from <http://www.scq.ubc.ca>

<https://www.youtube.com/watch?v=9kOGOY7vthk>

# Amino Acid

The side chain (R) determines the type of the amino acid

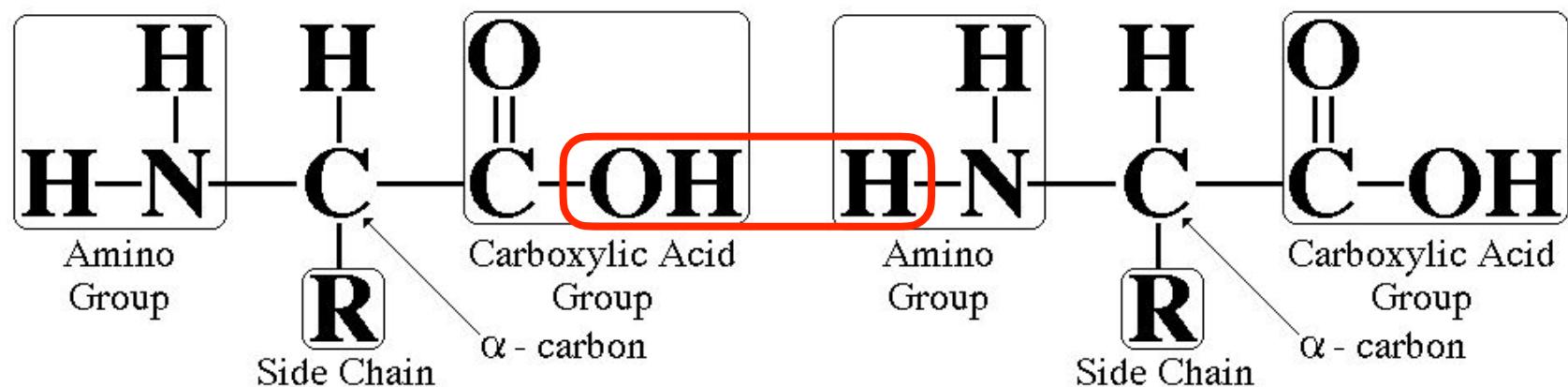


# Physico-chemical Properties

The properties of the amino acid depends on the side chain

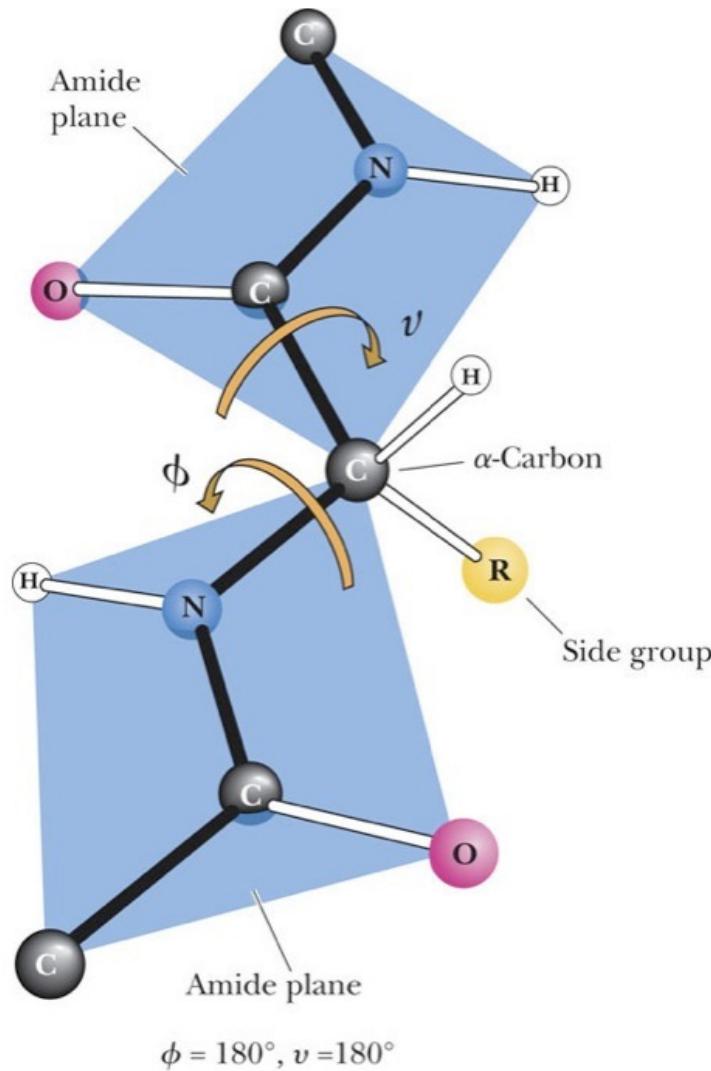
Amino acid	Abbrev.	Side chain	Hydro-phobic	Polar	Charged	Small	Tiny	Aromatic or Aliphatic	van der Waals volume	Codon	Occurrence in proteins (%)
Alanine	Ala, A	-CH <sub>3</sub>	X	-	-	X	X	-	67	GCU, GCC, GCA, GCG	7.8
Cysteine	Cys, C	-CH <sub>2</sub> SH	X	-	-	X	-	-	86	UGU, UGC	1.9
Aspartate	Asp, D	-CH <sub>2</sub> COOH	-	X	negative	X	-	-	91	GAU, GAC	5.3
Glutamate	Glu, E	-CH <sub>2</sub> CH <sub>2</sub> COOH	-	X	negative	-	-	-	109	GAA, GAG	6.3
Phenylalanine	Phe, F	-CH <sub>2</sub> C <sub>6</sub> H <sub>5</sub>	X	-	-	-	-	Aromatic	136	UUU, UUC	3.9
Glycine	Gly, G	-H	X	-	-	X	X	-	48	GGU, GGC, GGA, GGG	7.2
Histidine	His, H	-CH <sub>2</sub> C <sub>3</sub> H <sub>3</sub> N <sub>2</sub>	-	X	positive	-	-	Aromatic	118	CAU, CAC	2.3
Isoleucine	Ile, I	-CH(CH <sub>3</sub> )CH <sub>2</sub> CH <sub>3</sub>	X	-	-	-	-	Aliphatic	124	AUU, AUC, AUA	5.3
Lysine	Lys, K	-(CH <sub>2</sub> ) <sub>4</sub> NH <sub>2</sub>	-	X	positive	-	-	-	135	AAA, AAG	5.9
Leucine	Leu, L	-CH <sub>2</sub> CH(CH <sub>3</sub> ) <sub>2</sub>	X	-	-	-	-	Aliphatic	124	UUA, UUG, CUU, CUC, CUA, CUG	9.1
Methionine	Met, M	-CH <sub>2</sub> CH <sub>2</sub> SCH <sub>3</sub>	X	-	-	-	-	-	124	AUG	2.3
Asparagine	Asn, N	-CH <sub>2</sub> CONH <sub>2</sub>	-	X	-	X	-	-	96	AAU, AAC	4.3
Proline	Pro, P	-CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> -	X	-	-	X	-	-	90	CCU, CCC, CCA, CCG	5.2
Glutamine	Gln, Q	-CH <sub>2</sub> CH <sub>2</sub> CONH <sub>2</sub>	-	X	-	-	-	-	114	CAA, CAG	4.2
Arginine	Arg, R	-(CH <sub>2</sub> ) <sub>3</sub> NH-C(NH)NH <sub>2</sub>	-	X	positive	-	-	-	148	CGU, CGC, CGA, CGG, AGA, AGG	5.1
Serine	Ser, S	-CH <sub>2</sub> OH	-	X	-	X	X	-	73	UCU, UCC, UCA, UCG, AGU, AGC	6.8
Threonine	Thr, T	-CH(OH)CH <sub>3</sub>	X	X	-	X	-	-	93	ACU, ACC, ACA, ACG	5.9
Valine	Val, V	-CH(CH <sub>3</sub> ) <sub>2</sub>	X	-	-	X	-	Aliphatic	105	GUU, GUC, GUA, GUG	6.6
Tryptophan	Trp, W	-CH <sub>2</sub> C <sub>6</sub> H <sub>5</sub> N	X	-	-	-	-	Aromatic	163	UGG	1.4
Tyrosine	Tyr, Y	-CH <sub>2</sub> C <sub>6</sub> H <sub>4</sub> OH	X	X	-	-	-	Aromatic	141	UAU, UAC	3.2

# Peptide Bond



# Torsion Angles

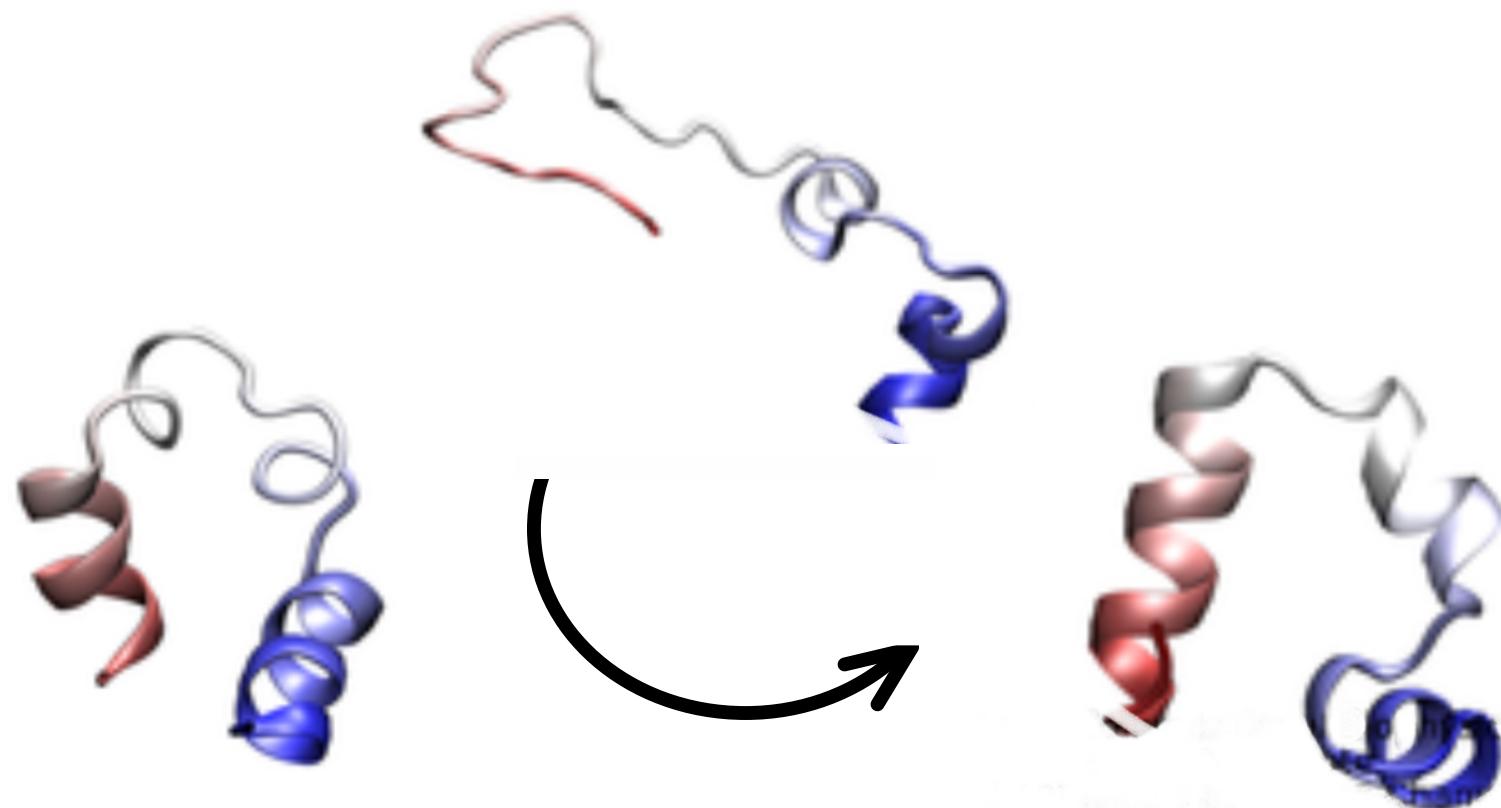
Backbone torsion angles determine the structure of the protein



# Protein folding

Protein folding is the process by which a protein assumes its native structure from the unfolded structure

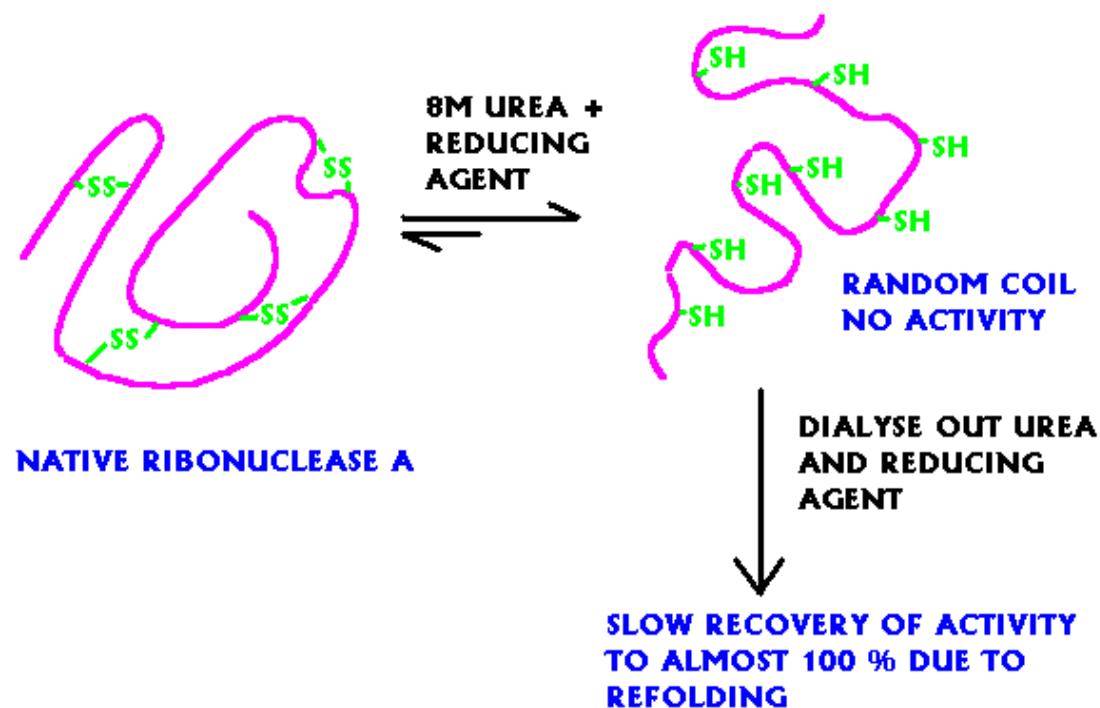
T T C C P S I V A R S N F N V C R L P G T P E A L C A T  
Y T G C I I I P G A T C P G D Y A N



# The Anfinsen's hypothesis

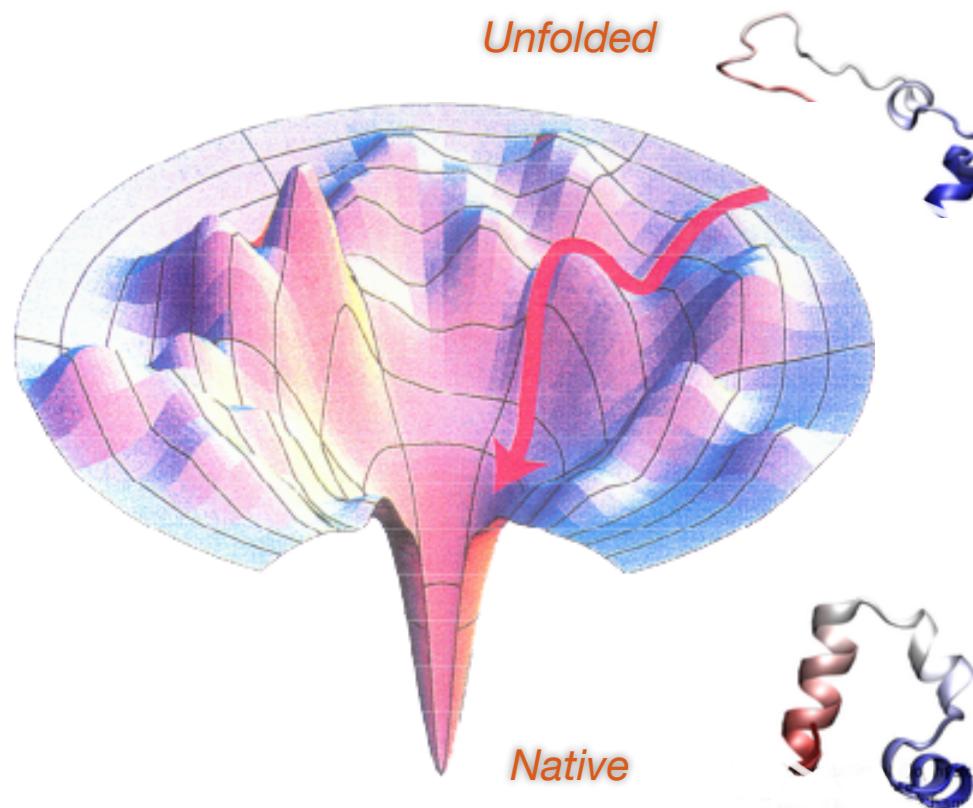
The sequence contains all the information to specify 3-D structure

Anfinsen showed that denatured ribonuclease A could be re-activated removing the denaturant.



# Levinthal's paradox

A protein chain composed by 100 residues with 2 possible conformations has  $2^{100}$  ( $\sim 10^{30}$ ) possible conformations. Considering a time-step of  $10^{-12}$  s for visiting each conformation, the folding process would take  $10^{18}$  s, that is longer than the age of our Universe ( $2\text{-}3 \times 10^{17}$ s)



# The Anfinsen's Dogma

**Uniqueness:** requires that the sequence does **not have any other configuration with a comparable free energy.**

**Stability:** **small changes** in the surrounding environment **not affect the structure of the stable conformation.** This can be pictured as a free energy surface that looks more like a funnel and the free energy surface around the native state must be rather steep and high, in order to provide stability.

**Kinetic accessibility:** means that the path in the **free energy surface** from the unfolded to the folded state **must be reasonably smooth** or, in other words, that the folding of the chain must not involve highly complex changes in the shape.

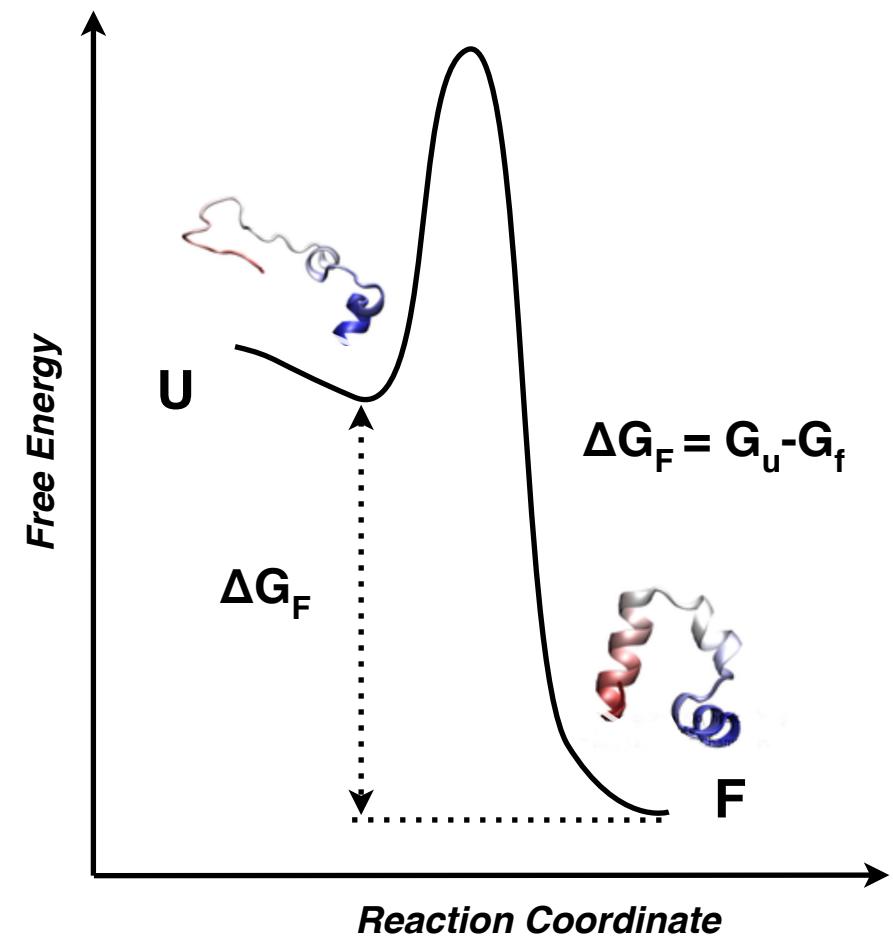
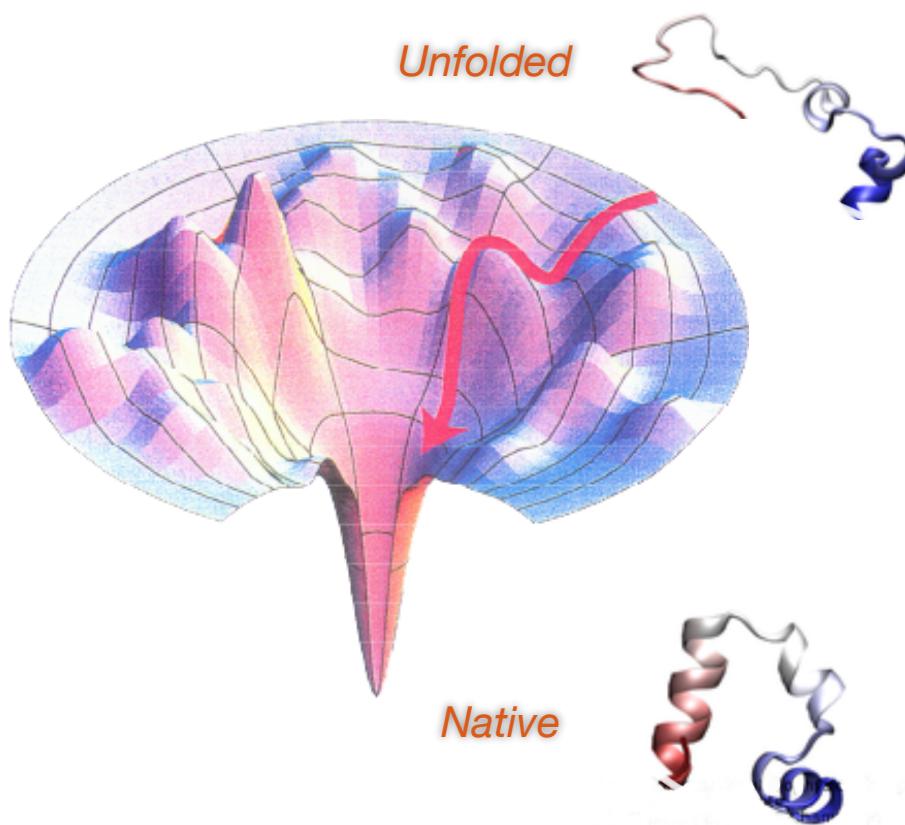
# Aspects of the same problem

The solution of the protein folding consists in the understanding of three different aspects of the problem:

- Estimate the **stability of the native conformation** and thermodynamic of the process.
- Define the mechanism and the **kinetic of the process**.
- Predict the native **three-dimensional structure** of the protein.

# Folding and stability

The folding free energy difference,  $\Delta G_F$ , is typically small, of the order of -5 to -15 kcal/mol for a globular protein (compared to e.g. -30 to -100 kcal/mol for a covalent bond).



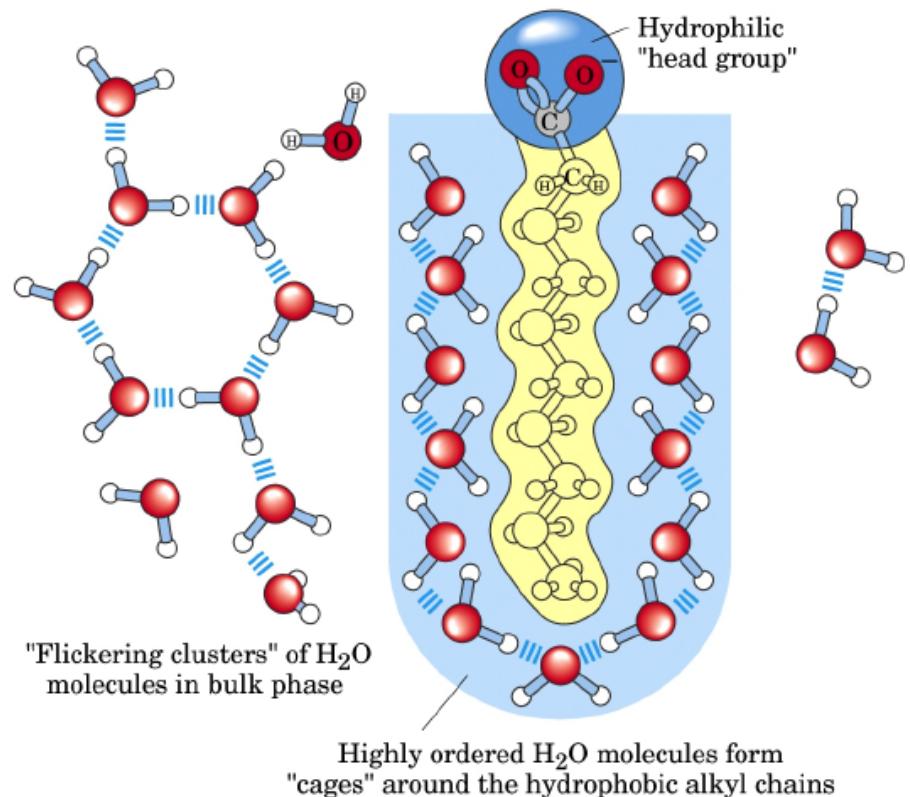
# Folding interactions

Several **electrostatic interactions** are contributing to the **stability** of the native state but they are **not the driving forces** in the folding process

Type	Examples	Binding energy (kcal/mol)	Change of free energy water to ethanol (kcal/mol)
Electrostatic interaction	Salt bridge $\text{---COO}^- \cdots \text{N}^+H_3\text{---}$	-5	-1
	Dipole-dipole $\begin{array}{c} \delta^+ \\ \backslash \\ \text{C=O} \\ / \end{array} \cdots \begin{array}{c} \delta^- \\ \backslash \\ \text{O=C} \\ / \end{array}$	+0.3	
Hydrogen bond	Water $\begin{array}{c} \text{H} \\   \\ \text{O---H} \cdots \text{O---H} \\   \\ \text{H} \end{array}$	-4	
	Protein backbone $\begin{array}{c} \backslash \\ \text{N---H} \cdots \text{O=C} \\ / \end{array}$	-3	
Dispersion forces	Aliphatic hydrogen $\begin{array}{c}   \\ \text{---C---H} \cdots \text{H---C---} \\   \end{array}$	-0.03	
Hydrophobic forces	Side chain of Phe		-2.4

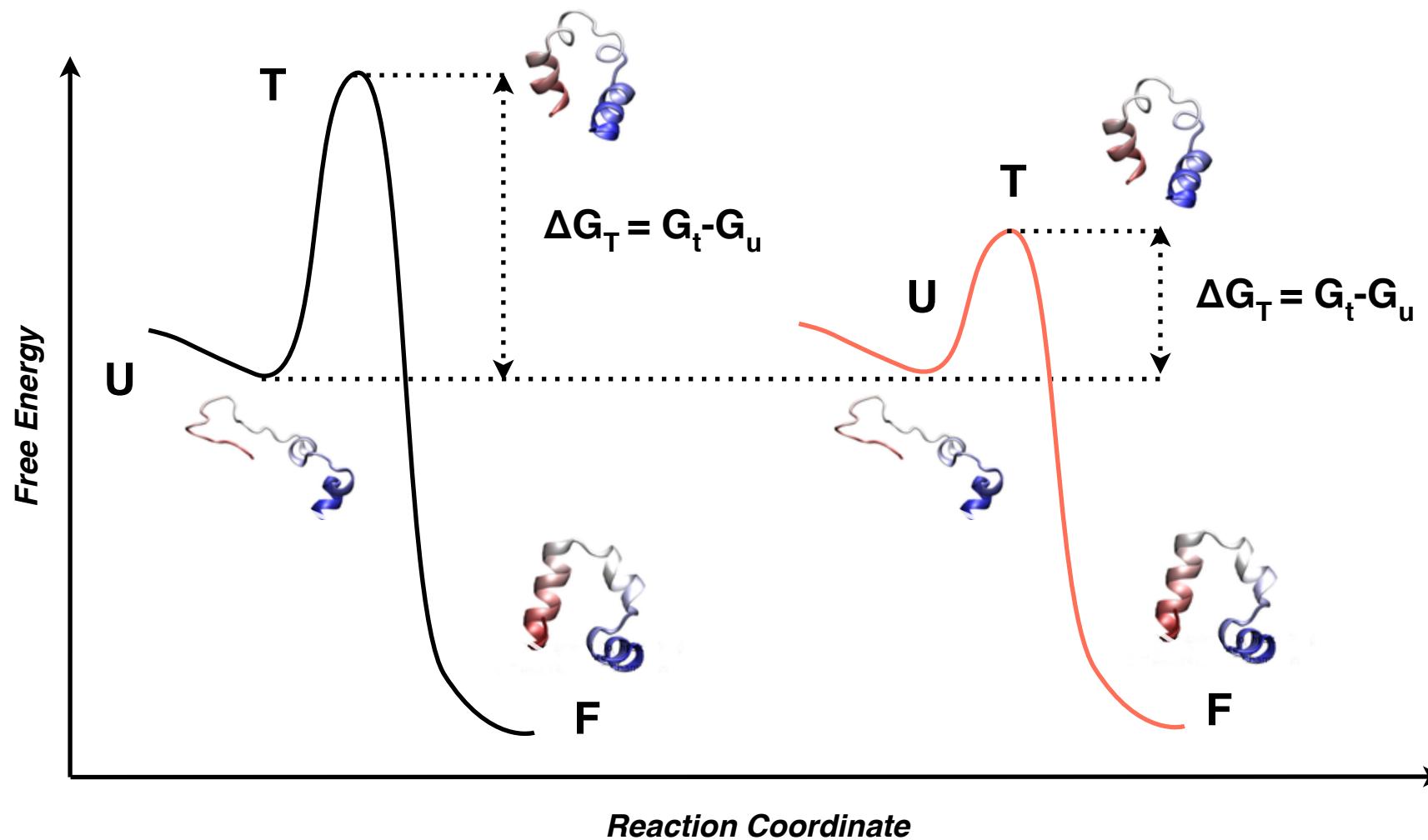
# Hydrophobic effect

- Water molecules form a cage-like structure around the nonpolar molecule.
- The positive  $\Delta H$  is due to the fact that the cage has to be broken to transfer the nonpolar molecule.
- The positive  $\Delta S$  is due to the fact that the water molecules are less ordered (an increase in the degree of disorder) when the cage is broken.



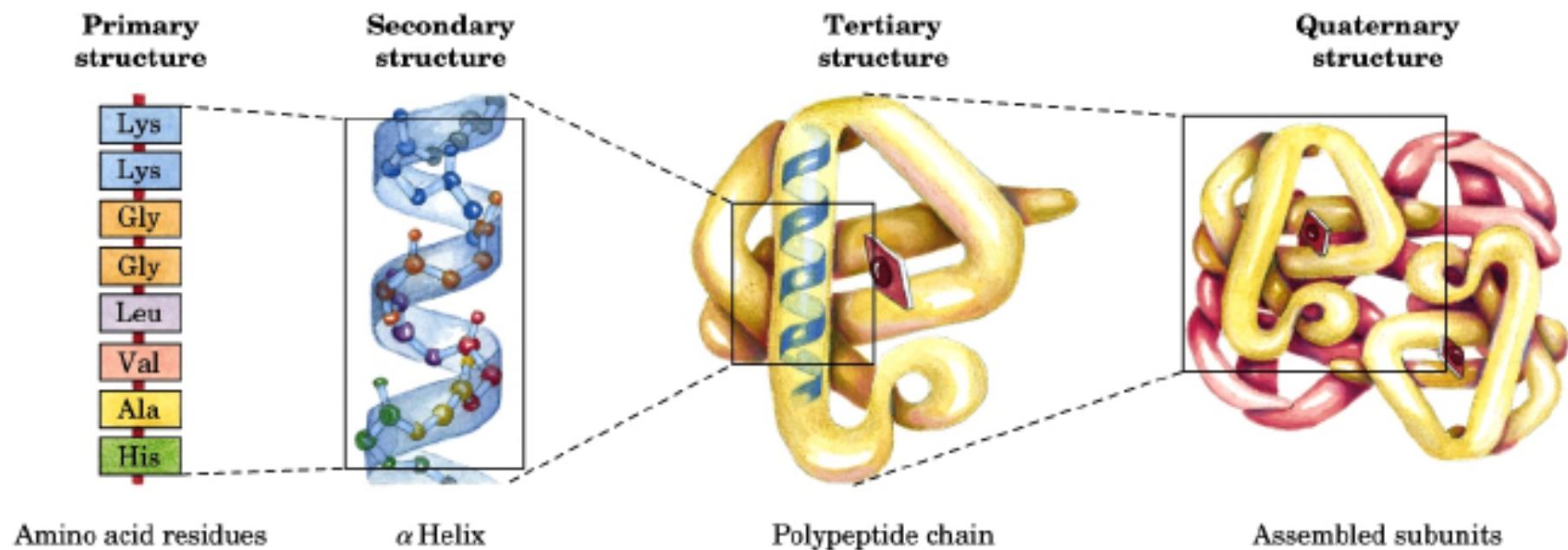
# Folding kinetics

The protein folding mechanism depends on the form of the free energy profile.  
Higher activation barrier corresponds to longer folding time



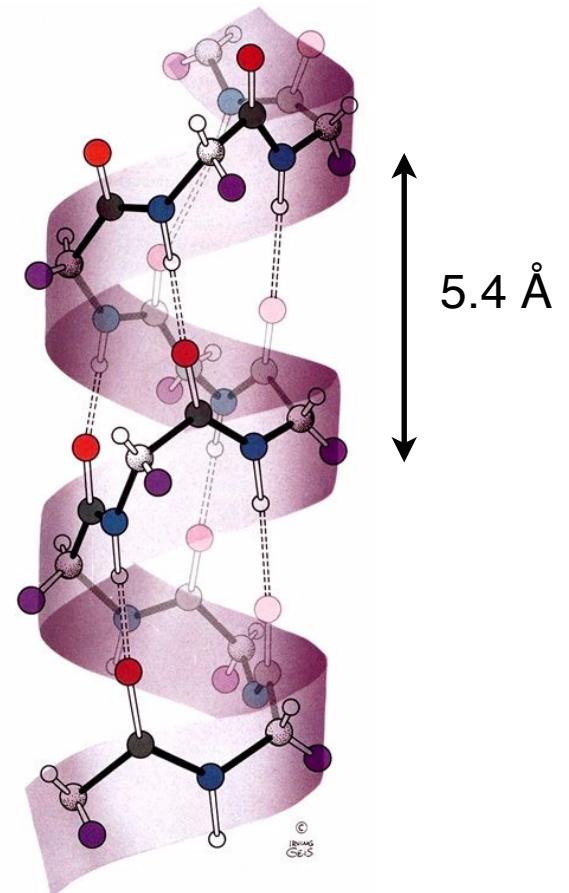
# Hierarchical organization of protein structure

Protein structure is defined by four levels of hierarchical organization.



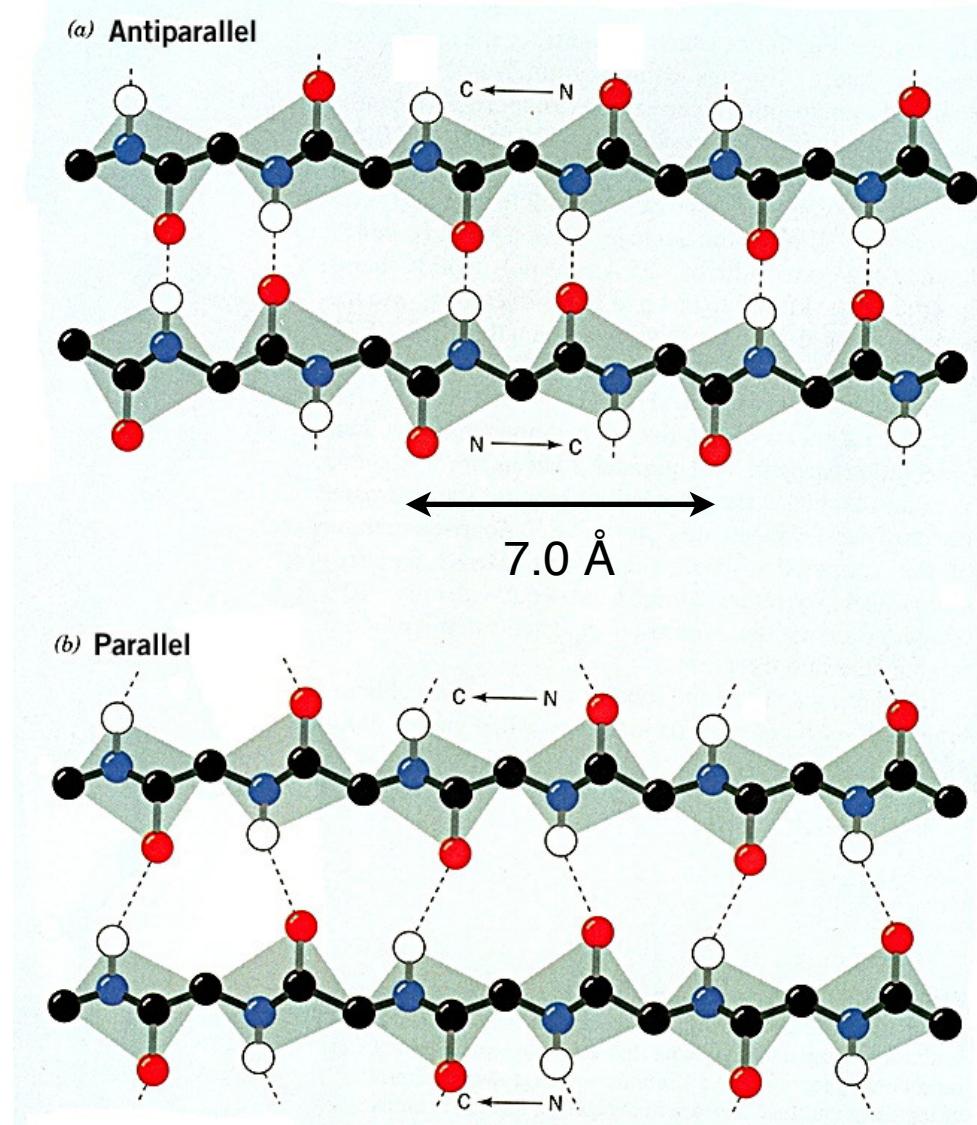
# Secondary structure (I)

- Helices observed in proteins are mostly right-handed.
- Typical  $\phi$ ,  $\psi$  values for residues in  $\alpha$ -helix are around  $-60^\circ$ ;  $-50^\circ$
- Side chains project backward and outward.
- The core of  $\alpha$ -helix is tightly packed.



# Secondary structure (II)

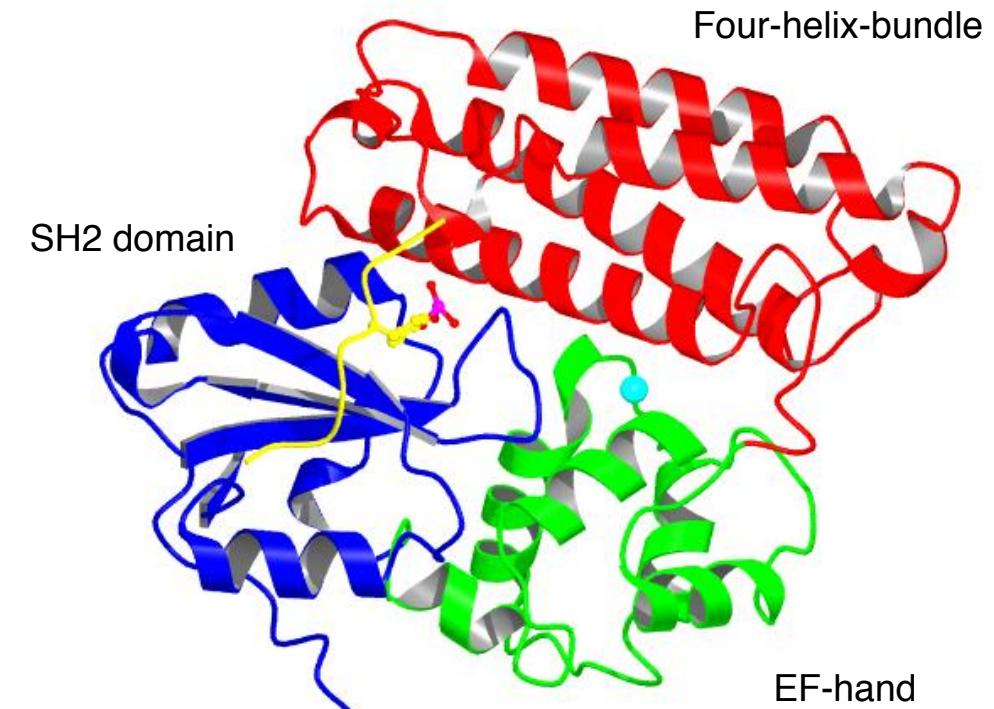
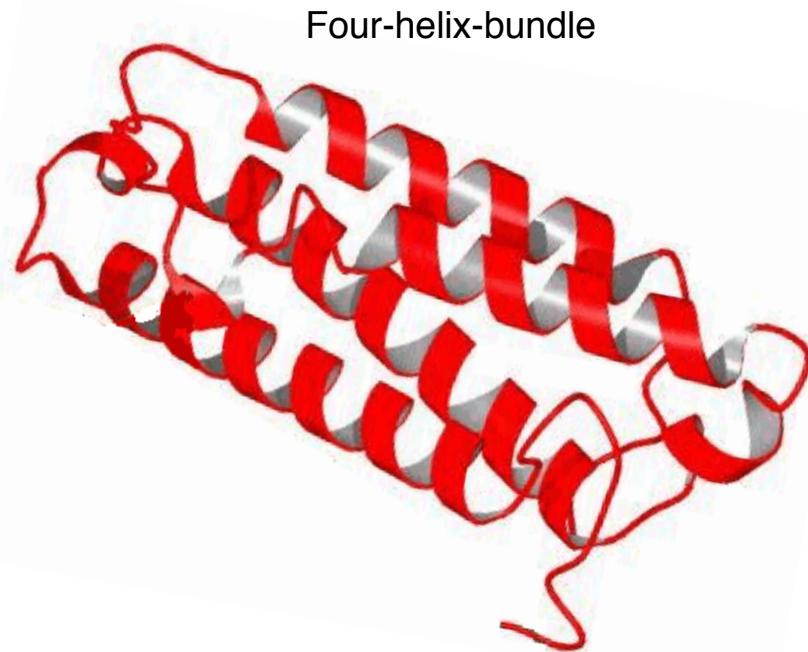
- Typical  $\phi$ ,  $\psi$  values for residues in  $\beta$ -sheet are around  $140^\circ$ ,  $-130^\circ$
- Side chains of neighboring residues project in opposite directions.
- The polypeptide is in a more extended conformation.
- Parallel  $\beta$ -sheets are less stable than anti-parallel  $\beta$ -sheets.



# More complex structures

The arrangements of secondary structural elements form the Tertiary Structure of the protein.

The complex of **two or more protein domains defines the Quaternary Structure**. In the example Four-helix-bundle, EF-hand and SH2 domains together form an integrated phosphoprotein that functions as a negative regulator of many signaling pathways from receptors at the cell surface.



# Protein at the NCBI

The Protein database is a collection of sequences from several resources accessible through Entrez

NCBI Resources How To

emidio.capriotti@hhu.de My NCBI Sign Out

Protein Protein Advanced Search Help

**Protein**

The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and function.

**Using Protein**

- [Quick Start Guide](#)
- [FAQ](#)
- [Help](#)
- [GenBank FTP](#)
- [RefSeq FTP](#)

**Protein Tools**

- [BLAST](#)
- [LinkOut](#)
- [E-Utilities](#)
- [Batch Entrez](#)

**Other Resources**

- [GenBank Home](#)
- [RefSeq Home](#)
- [CDD](#)
- [Structure](#)

<https://www.ncbi.nlm.nih.gov/protein/>

# Protein search

Using the name of the protein and the organism we can retrieve a specific protein

## P53 [Homo sapiens]

GenBank: BAC16799.1

[Identical Proteins](#) [FASTA](#) [Graphics](#)

---

[Go to:](#)

LOCUS	BAC16799	393 aa	linear	PRI 01-APR-2003
DEFINITION	P53 [Homo sapiens].			
ACCESSION	BAC16799			
VERSION	BAC16799.1			
DBSOURCE	accession <a href="#">AB082923.1</a>			
KEYWORDS	.			
SOURCE	Homo sapiens (human)			
ORGANISM	<u>Homo sapiens</u> Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.			
REFERENCE	1			
AUTHORS	Azuma,K., Shichijo,S., Maeda,Y., Nakatsura,T., Nonaka,Y., Fujii,T., Koike,K. and Itoh,K.			
TITLE	Mutated p53 gene encodes a nonmutated epitope recognized by HLA-B*4601-restricted and tumor cell-reactive CTLs at tumor site			
JOURNAL	Cancer Res. 63 (4), 854-858 (2003)			
PUBMED	<a href="#">12591737</a>			
REFERENCE	2 (residues 1 to 393)			
AUTHORS	Shichijo,S. and Itoh,K.			
TITLE	Direct Submission			
JOURNAL	Submitted (26-MAR-2002) Shigeki Shichijo, Kurume Univ. School of Med., Dep. Immunol.; 67-Asahi-machi, Kurume, Fukuoka 830-0011, Japan (E-mail:shichijo@med.kurume-u.ac.jp, Tel:81-942-31-7551, Fax:81-942-31-7699)			
FEATURES	Location/Qualifiers			

# Protein Sequence DB

The main database of protein sequences is UniProt which is composed by SwissProt and TrEMBL

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.

**UniProtKB**  
UniProt Knowledgebase  
Swiss-Prot (563,552)  
Manually annotated and reviewed.  
Records with information extracted from literature and curator-evaluated computational analysis.

**UniRef**  
The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records.

**UniParc**  
UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world.

**Proteomes**  
A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes.

**Supporting data**

- Literature citations
- Cross-ref. databases
- Taxonomy
- Diseases
- Subcellular locations
- Keywords

<https://www.uniprot.org>

# UniProt Composition

Database of annotated proteins

- Swiss-Prot: Manually annotated ~560K
- TrEMBL: Automatically annotated ~195M

## UniProtKB

UniProt Knowledgebase

Swiss-Prot (563,552)



Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (195,104,019)



Automatically annotated and not reviewed.

Records that await full manual annotation.

# The SwissProt

SwissProt contains all the **proteins that have been manually annotated** using information extracted from literature.

## UniProtKB/Swiss-Prot

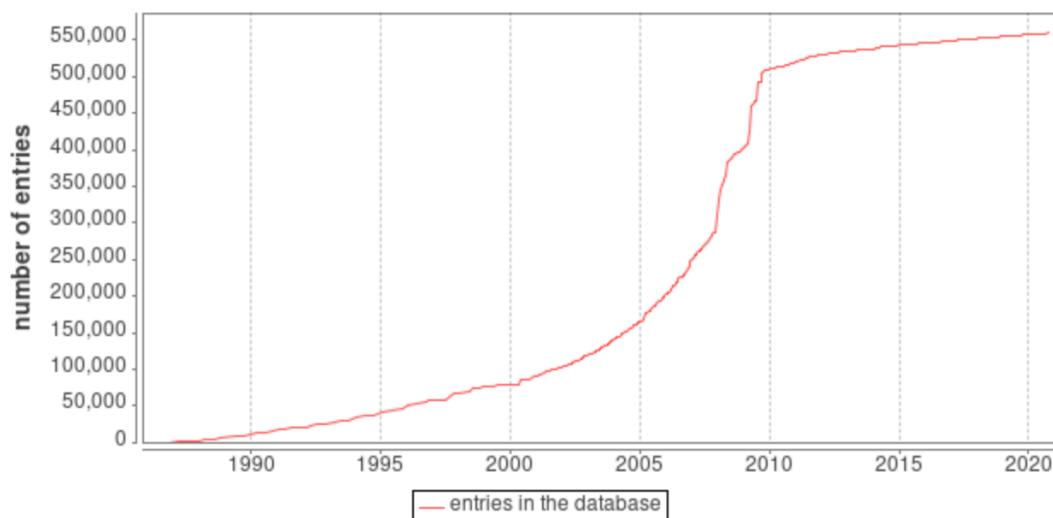


UniProtKB/Swiss-Prot is the expertly curated component of UniProtKB (produced by the UniProt consortium). It contains hundreds of thousands of protein descriptions, including function, domain structure, subcellular location, post-translational modifications and functionally characterized variants.

💡 UniProt is one of the most widely used protein information resources in the world, and an ELIXIR Core Data Resource.

[Browse the resource website](#)

### Number of entries in UniProtKB/Swiss-Prot over time



What you can do with this resource?

[Query and retrieval](#),  
[Protein feature detection](#)

Browse these keywords in Expasy

[Protein sites, features and motifs](#),  
[Data submission, annotation and curation](#),  
[Proteomics](#),  
[Molecular interactions, pathways and networks](#),  
[Function analysis](#),  
[Protein structure analysis](#),  
[Taxonomy](#),  
[Keyword](#),  
[UniProt accession](#),  
[Protein sequence record](#),  
[Text](#)

# The function

Acts as a tumor suppressor in many tumor types; induces growth arrest or apoptosis depending on the physiological circumstances and cell type.  
11 publications

UniProt

UniProtKB

Advanced Search

BLAST Align Retrieve/ID mapping Help Contact

P04637 - P53\_HUMAN

Basket

Protein: Cellular tumor antigen p53  
Gene: TP53  
Organism: Homo sapiens (Human)  
Status: Reviewed - Annotation score: 00000 - Experimental evidence at protein level<sup>1</sup>

Display: None

Function: Acts as a tumor suppressor in many tumor types; induces growth arrest or apoptosis depending on the physiological circumstances and cell type. Involved in cell cycle regulation as a trans-activator that acts to negatively regulate cell division by controlling a set of genes required for this process. One of the activated genes is an inhibitor of cyclin-dependent kinases. Apoptosis induction seems to be mediated either by stimulation of BAX and FAS antigen expression, or by repression of Bcl-2 expression. In cooperation with mitochondrial PP1F is involved in activating oxidative stress-induced necrosis; the function is largely independent of transcription. Induces the transcription of long intergenic non-coding RNA p21 (lncRNA-p21) and lncRNA-Mkln1. LncRNA-p21 participates in TP53-dependent transcriptional repression leading to apoptosis and seem to have effect on cell-cycle regulation. Implicated in Notch signaling cross-over. Prevents CDK7 kinase activity when associated to CAK complex in response to DNA damage, thus stopping cell cycle progression. Isoform 2 enhances the transactivation activity of isoform 1 from some but not all TP53-inducible promoters. Isoform 4 suppresses transactivation activity and impairs growth suppression mediated by isoform 1. Isoform 7 inhibits isoform 1-mediated apoptosis. [11 Publications](#)

Cofactor<sup>1</sup>: Zn<sup>2+</sup>  
Note: Binds 1 zinc ion per subunit.

Sites:

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Site <sup>1</sup>	120 – 120		1 Interaction with DNA			
Metal binding <sup>1</sup>	176 – 176		1 Zinc			
Metal binding <sup>1</sup>	179 – 179		1 Zinc			
Metal binding <sup>1</sup>	238 – 238		1 Zinc			
Metal binding <sup>1</sup>	242 – 242		1 Zinc			

Regions:

# Getting the information

The SwissProt **fasta file contains all the sequences** in the database and the **text file contains** all the information including **annotation**.

The fasta and text files can be downloaded using the following links

[http://www.uniprot.org/uniprot/P53\\_HUMAN.fasta](http://www.uniprot.org/uniprot/P53_HUMAN.fasta)

[http://www.uniprot.org/uniprot/P53\\_HUMAN.txt](http://www.uniprot.org/uniprot/P53_HUMAN.txt)

More complex queries:

[http://www.uniprot.org/help/programmatic\\_access](http://www.uniprot.org/help/programmatic_access)

```
ID  P53_HUMAN          Reviewed;      393 AA.
AC  P04637; Q15086; Q15087; Q15088; Q16535; Q16807; Q16808; Q16809;
AC  Q16810; Q16811; Q16848; Q2XN98; Q3LRW1; Q3LRW2; Q3LRW3; Q3LRW4;
AC  Q3LRW5; Q86UG1; Q8J016; Q99659; Q9BTM4; Q9HAQ8; Q9NP68; Q9NPJ2;
AC  Q9NZD0; Q9UBI2; Q9UQ61;
DT  13-AUG-1987, integrated into UniProtKB/Swiss-Prot.
DT  24-NOV-2009, sequence version 4.
DT  04-FEB-2015, entry version 228.
DE  RecName: Full=Cellular tumor antigen p53;
DE  AltName: Full=Antigen NY-CO-13;
DE  AltName: Full=Phosphoprotein p53;
DE  AltName: Full=Tumor suppressor p53;
GN  Name=TP53; Synonyms=P53;
OS  Homo sapiens (Human).
OC  Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC  Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC  Catarrhini; Hominidae; Homo.
OX  NCBI_TaxID=9606;
RN  [1]
RP  NUCLEOTIDE SEQUENCE [mRNA] (ISOFORM 1).
RX  PubMed=4006916;
RA  Zakut-Houri R., Bienz-Tadmor B., Givol D., Oren M. ;
RT  "Human p53 cellular tumor antigen: cDNA sequence and expression in COS
RT  cells." ;
RL  EMBO J. 4:1251-1255(1985).
```

# Exercise

From the UniProt FTP web site (<ftp://ftp.expasy.org/databases/uniprot/>) download the Human protein UP000005640\_9606 in fasta format.

- What is the total number of human proteins in the SwissProt and TrEMBL dataset?
- Given the fasta file containing the protein sequence of P53 what is the total number of residues?

# The Protein Data Bank

The largest repository of macromolecular structures obtained mainly by X-ray crystallography and NMR

The screenshot shows the RCSB PDB homepage. At the top, a dark blue header bar contains navigation links: Deposit, Search, Visualize, Analyze, Download, Learn, More, and MyPDB. Below the header is the RCSB PDB logo and a banner stating "170383 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education". A search bar with the placeholder "Enter search term(s)" and a magnifying glass icon is positioned next to the banner. Below the search bar are links for "Advanced Search" and "Browse Annotations". To the right of the search bar is a large, semi-transparent image of a molecular structure. At the bottom of the header is a row of logos for PDB-101, Worldwide Protein Data Bank, EMDDataResource, NDB, and the Worldwide Protein Data Bank Foundation. On the left side, a vertical sidebar menu lists "Welcome", "Deposit", "Search", "Visualize", "Analyze", "Download", and "Learn", each with a corresponding icon. The main content area features a section titled "A Structural View of Biology" with text about the resource's purpose and its role in the wwPDB. It also includes a "COVID-19 CORONAVIRUS Resources" banner featuring a 3D model of a coronavirus. To the right, a section titled "October Molecule of the Month" displays a 3D molecular model of the Capsaicin Receptor TRPV1, colored in green and blue.

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB

170383 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Enter search term(s)

Advanced Search | Browse Annotations

PDB-101 Worldwide Protein Data Bank EMDDataResource NDB NUCLEIC ACID DATABASE Worldwide Protein Data Bank Foundation

Welcome

Deposit

Search

Visualize

Analyze

Download

Learn

A Structural View of Biology

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data. The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

COVID-19 CORONAVIRUS Resources

October Molecule of the Month

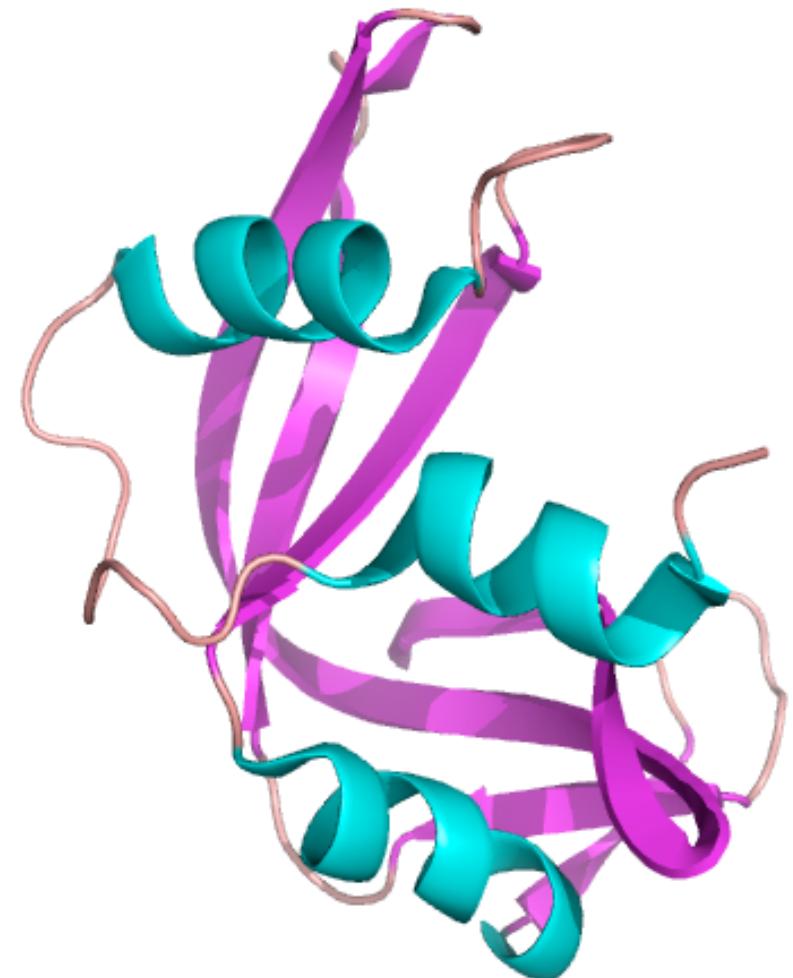
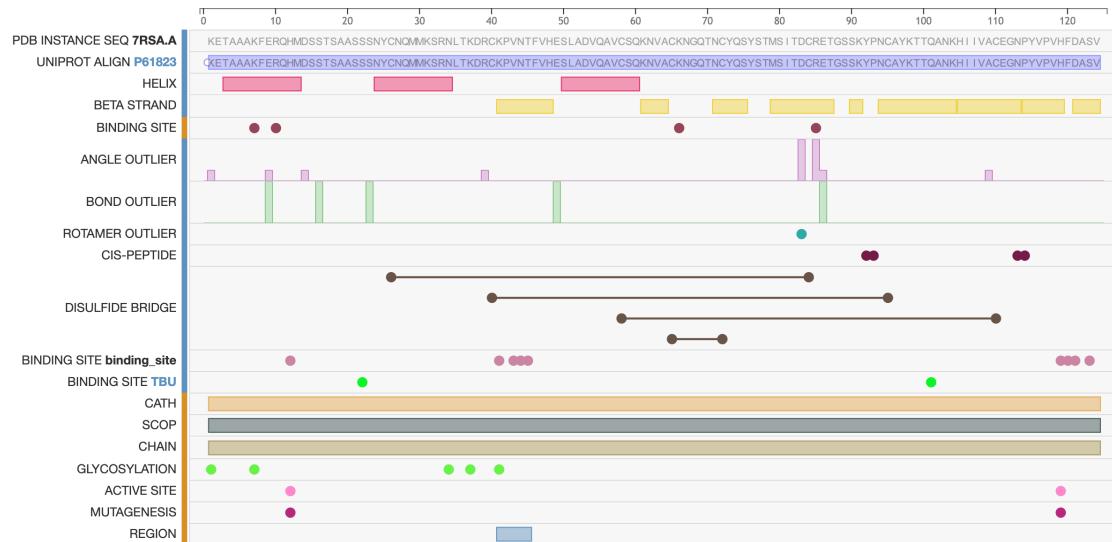
Capsaicin Receptor TRPV1

<http://rcsb.org>

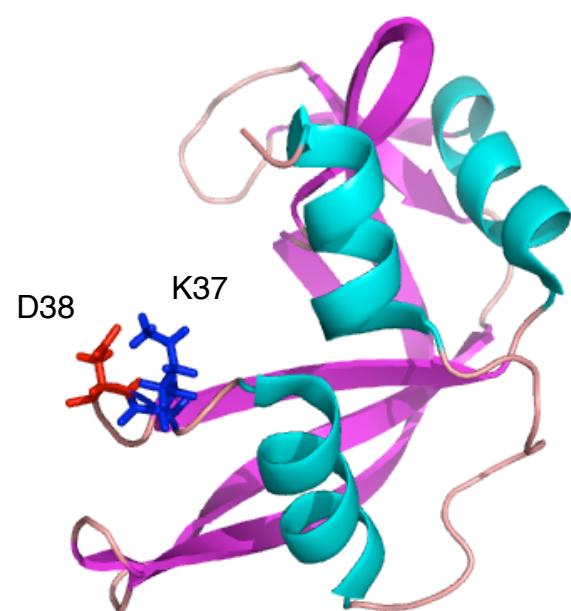
<http://ftp.rcsb.org/pub/pdb/>

# The Bovine Ribonuclease A

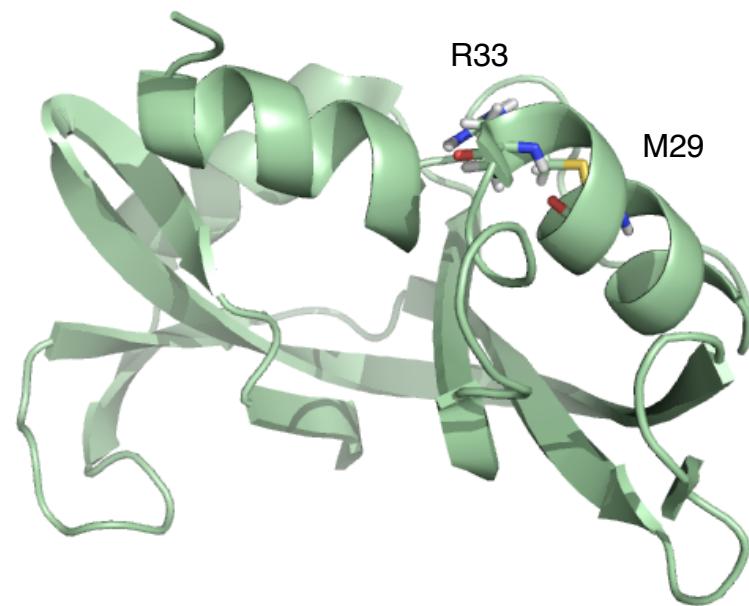
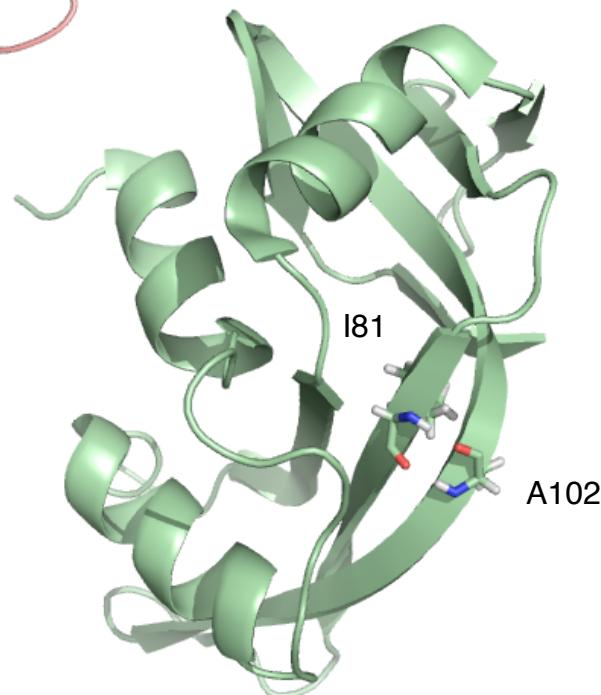
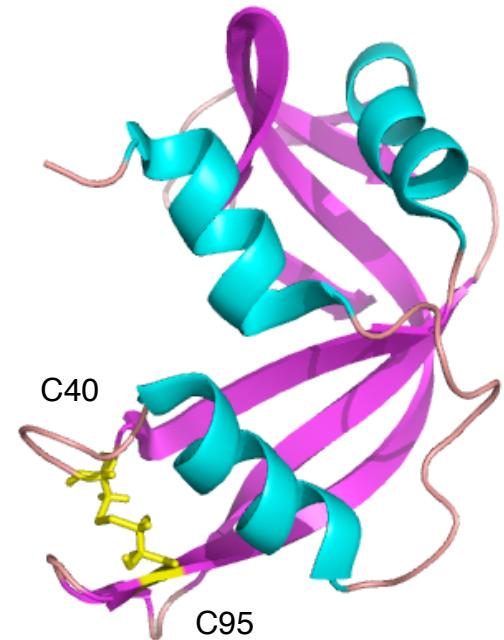
Ribonuclease A (RNase A) is a [pancreatic ribonuclease](#) that cleaves single-stranded RNA.



# Bonds and interactions



Examples of salt bridge, disulfide bond and hydrogen bonds in ribonuclease A



# PDB File

The most important information are the atomic coordinates.

		AT	RES	CH	POS	X	Y	Z	
ATOM	2145	N	GLU	B	10	150.341	72.309	103.145	1.00 99.90 N
ATOM	2146	CA	GLU	B	10	150.096	71.519	101.907	1.00 99.90 C
ATOM	2147	C	GLU	B	10	150.425	70.046	102.190	1.00 99.90 C
ATOM	2148	O	GLU	B	10	151.326	69.770	102.983	1.00 99.90 O
ATOM	2149	CB	GLU	B	10	150.963	72.057	100.790	1.00 99.90 C
ATOM	2150	N	PRO	B	11	149.661	69.092	101.595	1.00 99.90 N
ATOM	2151	CA	PRO	B	11	149.856	67.644	101.778	1.00 99.90 C
ATOM	2152	C	PRO	B	11	150.783	66.845	100.844	1.00 99.90 C
ATOM	2153	O	PRO	B	11	151.938	66.593	101.185	1.00 99.90 O
ATOM	2154	CB	PRO	B	11	148.425	67.108	101.722	1.00 99.90 C
ATOM	2155	CG	PRO	B	11	147.816	67.948	100.672	1.00 99.90 C
ATOM	2156	CD	PRO	B	11	148.333	69.350	101.000	1.00 99.90 C
ATOM	2157	N	SER	B	12	150.258	66.422	99.691	1.00 99.90 N
ATOM	2158	CA	SER	B	12	150.965	65.585	98.710	1.00 99.90 C
ATOM	2159	C	SER	B	12	150.922	64.167	99.292	1.00 99.90 C
ATOM	2160	O	SER	B	12	150.493	63.222	98.632	1.00 99.90 O
ATOM	2161	CB	SER	B	12	152.410	66.042	98.440	1.00 99.90 C
ATOM	2162	OG	SER	B	12	152.907	65.499	97.219	1.00 99.90 O

# Exercise

Download the PDB file of the Ribonuclease A (PDB: 7RSA) from the web (<http://ftp.rcsb.org/pub/pdb/data/structures/all/pdb/pdb7rsa.ent.gz>) and perform the following tasks

- Run a shell command to calculate the number of residues of the protein?
- Write a python script to parse the PDB file.
- Modify the program to calculate the distance between two atoms and residues.
- Calculate the average and standard deviation of the distance between two consecutive  $\alpha$  carbons?

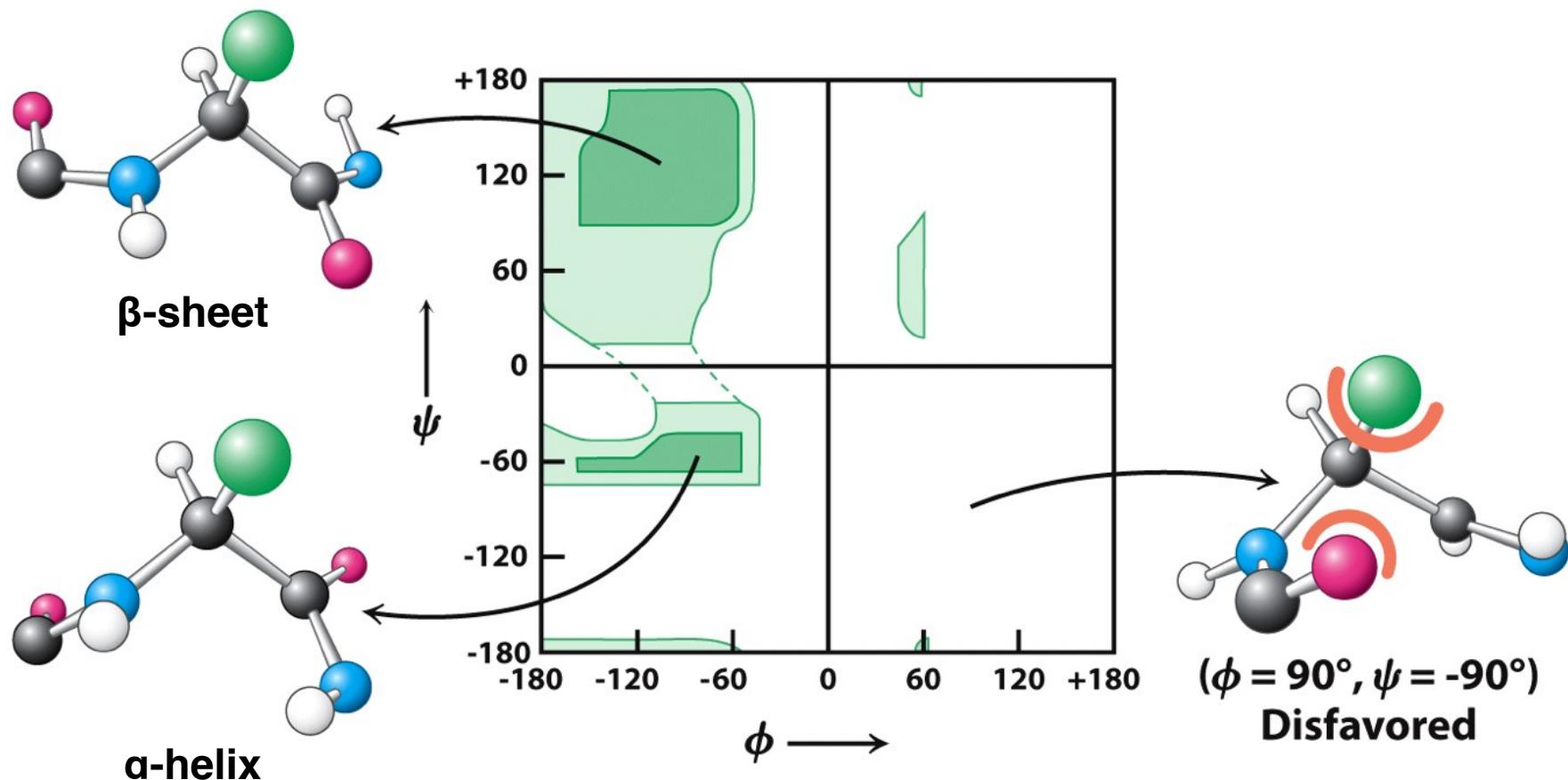
# Defining protein structure

Basic information for the characterization of the protein three-dimensional structures are:

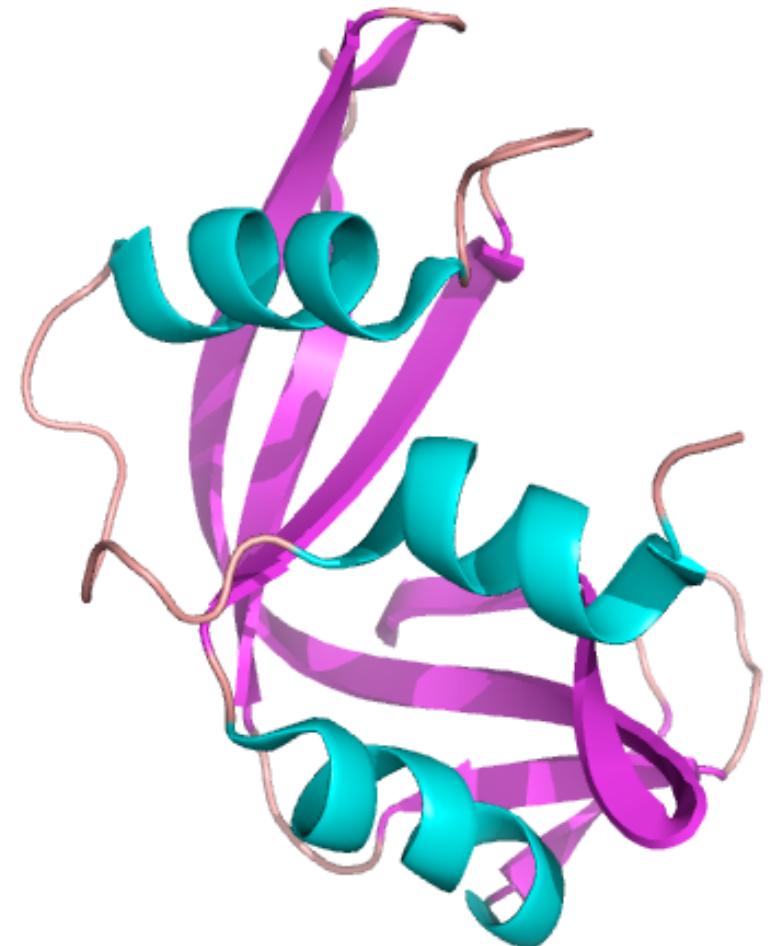
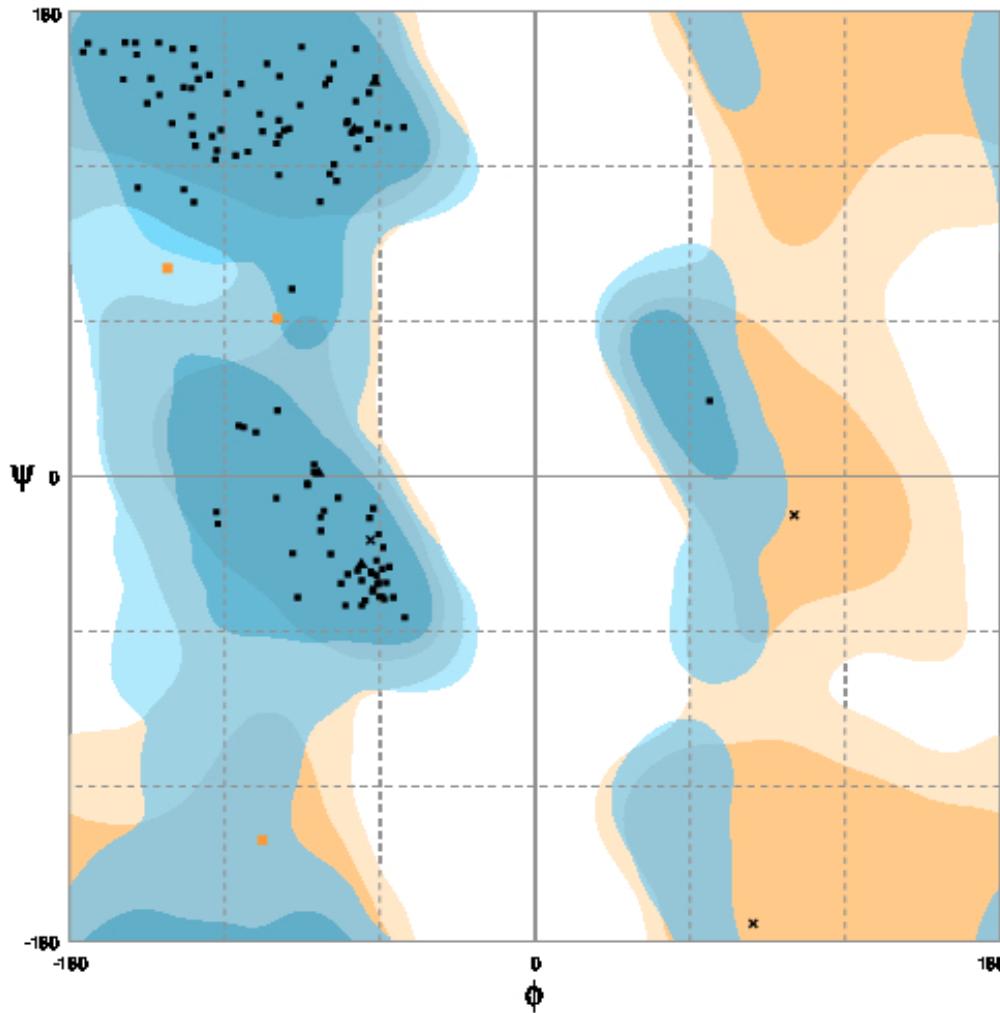
- $\phi, \psi$  values for each residue in the protein chain
- secondary structure
- solvent accessible area

# Ramachandran Plot

The backbone of the protein structure can be defined providing the list of  $\phi$ ,  $\psi$  angles for each residue in the chain.



# Ramachandran Analysis



<http://mordred.bioc.cam.ac.uk/~rapper/rampage.php>

# DSSP program

Program that implements the algorithm “Define Secondary Structure of Proteins”.

The method calculates different **features of the protein structure** such as the  $\phi$ ,  $\psi$  angles for each residue, its secondary structure and the solvent accessible area.

#	RESIDUE	AA	STRUCTURE	BP1	BP2	ACC	...	PHI	PSI	X-CA	Y-CA	Z-CA	
1	10	B	E			0	0	153	...	360.0	144.2	150.1	
2	11	B	P		+	0	0	83	...	-90.2	-84.0	149.9	
3	12	B	S	S	>>	S+	0	0	60	...	77.6	-51.1	151.0
4	13	B	A	T	34	S+	0	0	6	...	-82.3	73.7	151.3
5	14	B	D	T	3>	S+	0	0	39	...	-154.6	-41.3	147.5
6	15	B	W	H	<>	S+	0	0	170	...	-60.8	-41.6	148.0
7	16	B	L	H	X	S+	0	0	0	...	-62.9	-38.5	150.2
8	17	B	A	H	>	S+	0	0	3	...	-62.0	-58.1	147.4
9	18	B	T	H	X	S+	0	0	72	...	-56.4	-34.0	144.9

**SS**                    **SAA**                    **PHI**                    **PSI**

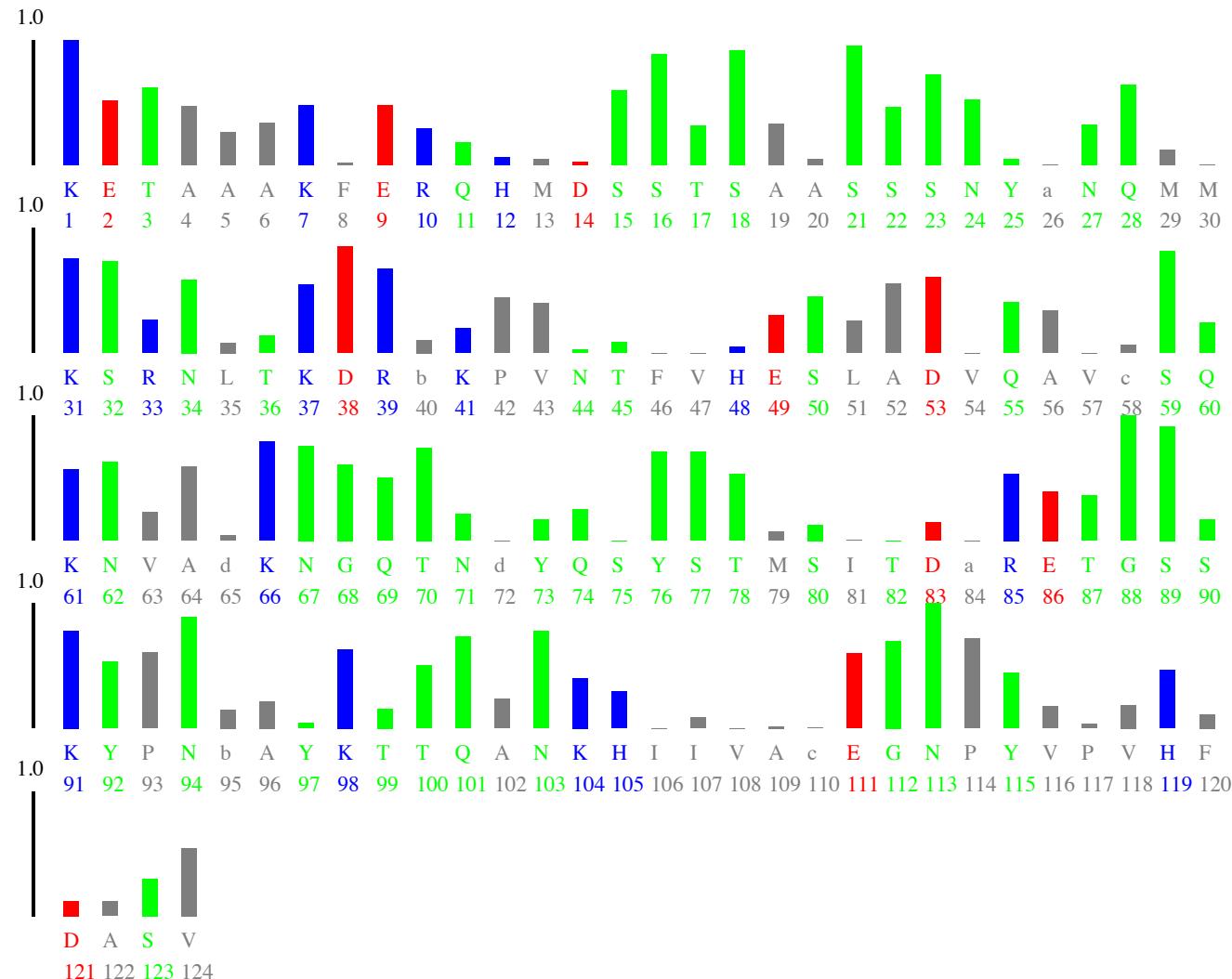
**cols=**            **17**                    **36-38**                    **104-109**                    **110-115**

DSSP: <ftp://ftp.cmbi.ru.nl/pub/software/dssp>  
more details at <https://swift.cmbi.umcn.nl/gv/dssp/>

Kabsch W, and Sander C, (1983). Biopolymers. 22 2577-2637.

# Relative solvent accessibility

The relative solvent accessible area is obtained dividing the accessible area of the residue by an estimation of the its maximum accessible surface.



# Exercise

Download the DSSP file of the Ribonuclease A (PDB: 7RSA) from the web (<ftp://ftp.cmbi.umcn.nl/pub/molbio/data/dssp/7rsa.dssp>) and answer the following questions

- What is the total number of residues in helical and extended conformations?
- What is the average value of the  $\phi$  and  $\psi$  angles for the residues in helical and extended conformations?
- Are the average values falling the the correct region of the Ramachandran plot?
- Considering the solvent accessibility values reported in the DSSP file, calculate the relative solvent accessible area for Lysine, Valine and Glutamine with maximum solvent accessible area of 205, 142 and 198 respectively
- Are this value compatible with the physico-chemical properties of the residues?