

# Protein Structure and Variants

**CB2-201 – Computational Biology and Bioinformatics**  
February 23, 2015

**Emidio Capriotti**  
<http://biifold.org/emidio>



**Biomolecules  
Folding and  
Disease**

Division of Informatics  
Department of Pathology

**UAB**  
THE UNIVERSITY OF  
ALABAMA AT BIRMINGHAM

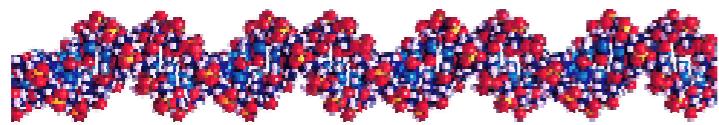


# Main data types

In molecular biology several type of data are available. Among the most common there are:

- **Sequences:** string representing the nucleotide and amino acid composition of DNA, RNA and protein.
- **Annotations:** collection of words with controlled vocabulary that describes property, function, and process in which a biomolecule is involved.
- **Structure:** 2D or 3D representation of a molecule describing how it is organized in the space.

# Molecular biology data



GenBank:

179,295,769

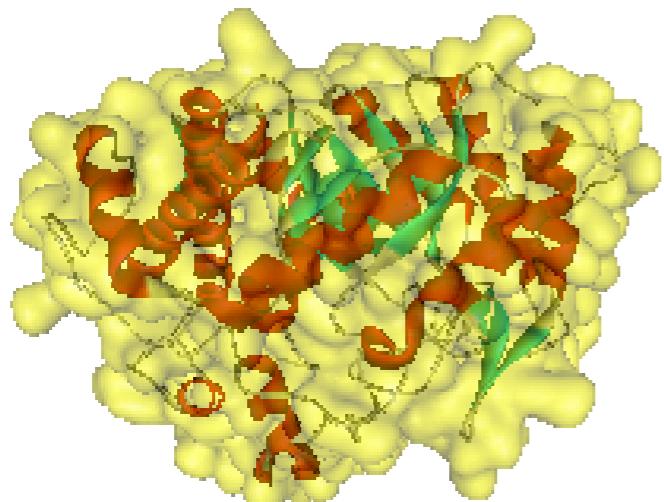
>BGAL\_SULSO BETA-GALACTOSIDASE *Sulfolobus solfataricus*.  
MYSFPNSFRFGWSQAGFQSEMGTGSEDPNTDWYKWHDPENMAAGLVSG  
DLPENGPGYWGNYKTFHDNAQKMGLKIAIRLNVEWSRIFPNPLPRPQNDFDE  
SKQDVTEVEINENELKRLDEYANKDALNHYREIFKDLKSRGLYFILNMYH  
WPLPLWLHDPIRVRRGDTGPGLSTRTVYEFARFSAYIAWKFDDLVDE  
YSTMNEPVVGGIGYVGVKSGFPPGYLSFELSRRHMYNIIQAHARAYDGI  
KSVSKKPVGIIYANSSFQPLTDKMEAVEMAENDNRWWFFDAIIRGEITR  
GNEKIVRDDLKGRLDWIGVNYYTRTVVKRTEKGYVSLGGYGHGCERNVS  
LAGLPTSDFGWEFFPEGLYDVLTKYWNRYHLYMYVTENGIADDADYQRPY  
YLVSHVYQVHRAINSGADVRGYLHWSDLADNEYEWASGFSMRFGLLKVDYNT  
KRLYWRPSALVYREIATNGAITDEIEHLNSVPPVKPLRH

UniRef90:

30,147,837

Swiss-Prot:

547,599



Protein Data Bank:

106,517

Protein:

98,954

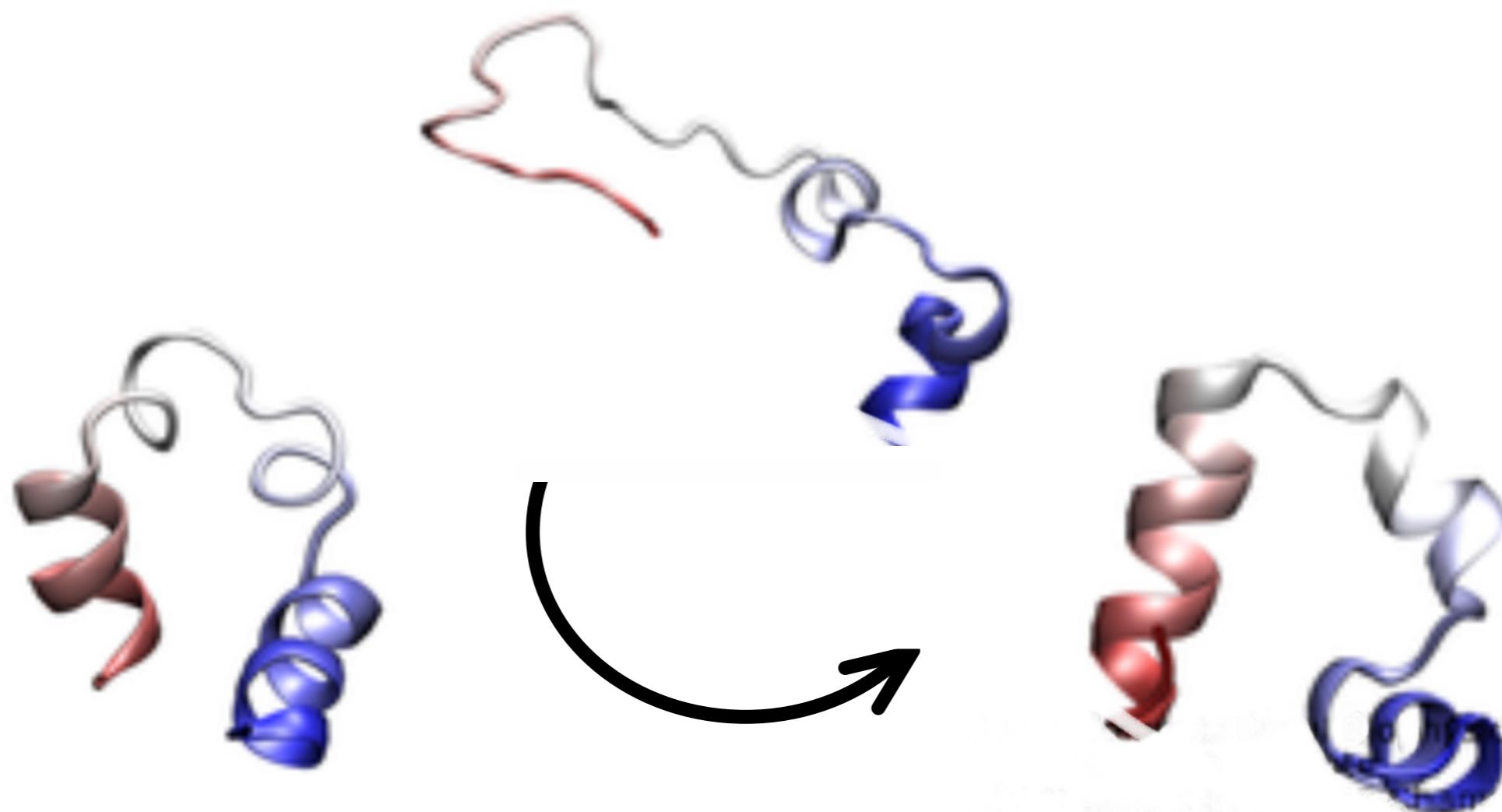
Nucleic Acids:

2,749

# Protein folding

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

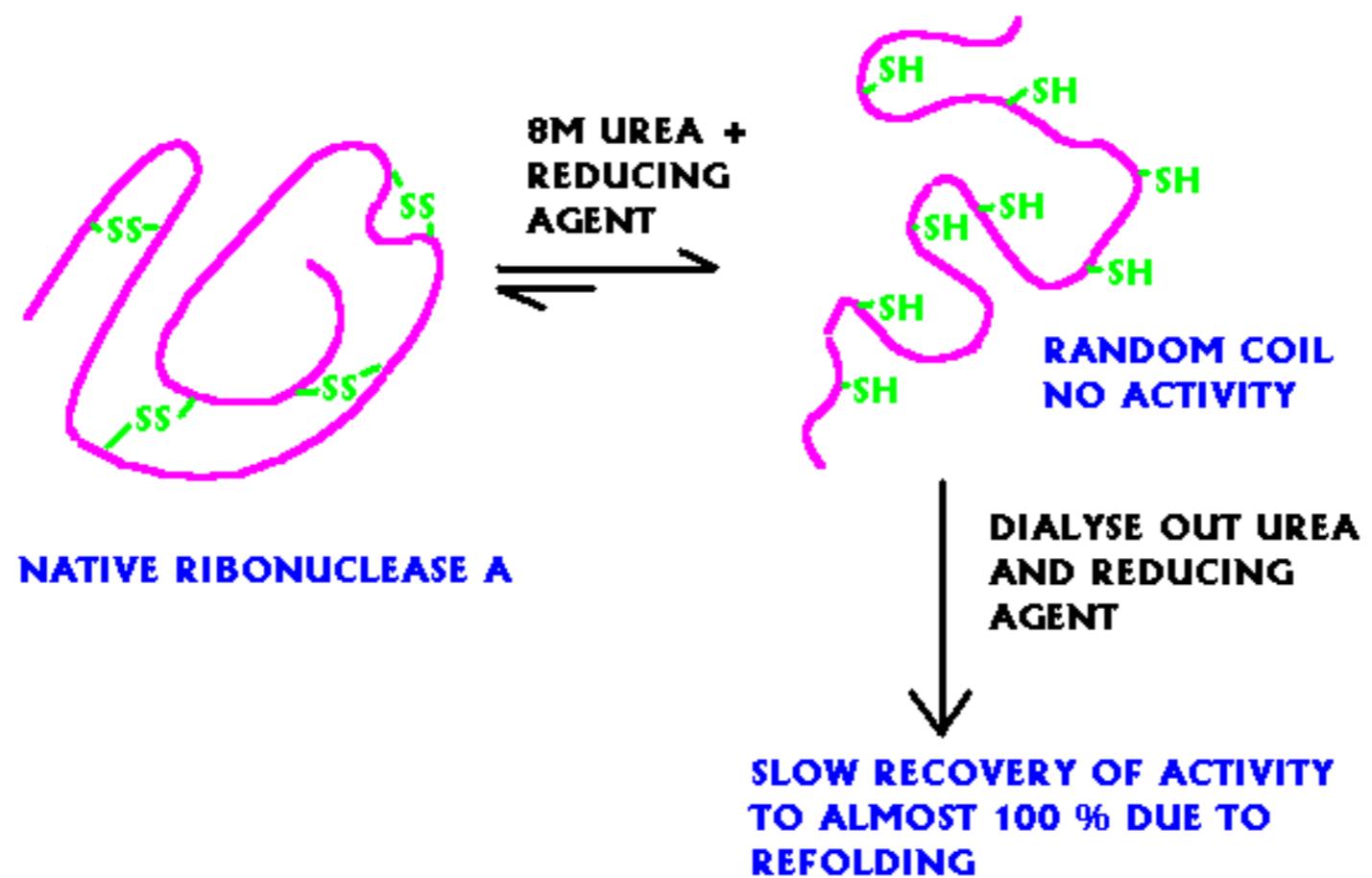
T T C C P S I V A R S N F N V C R L P G T P E A L C A T  
Y T G C I I I P G A T C P G D Y A N



# The Anfinsen's hypothesis

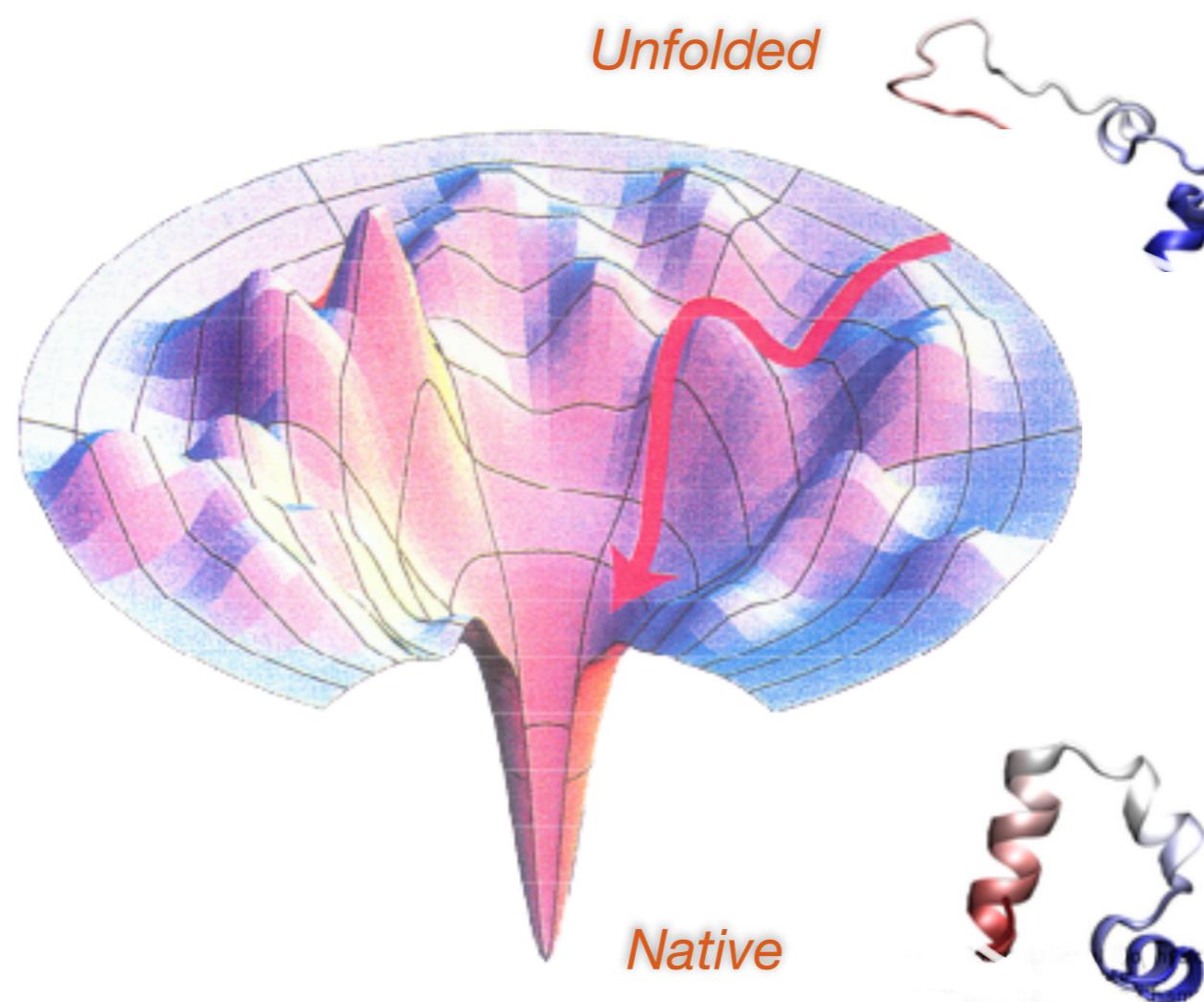
The sequence contains all the information to specify 3-D structure

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  $2^{100}$  ( $10^{30}$ ) 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 ( $2-3 \times 10^{17}$  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.

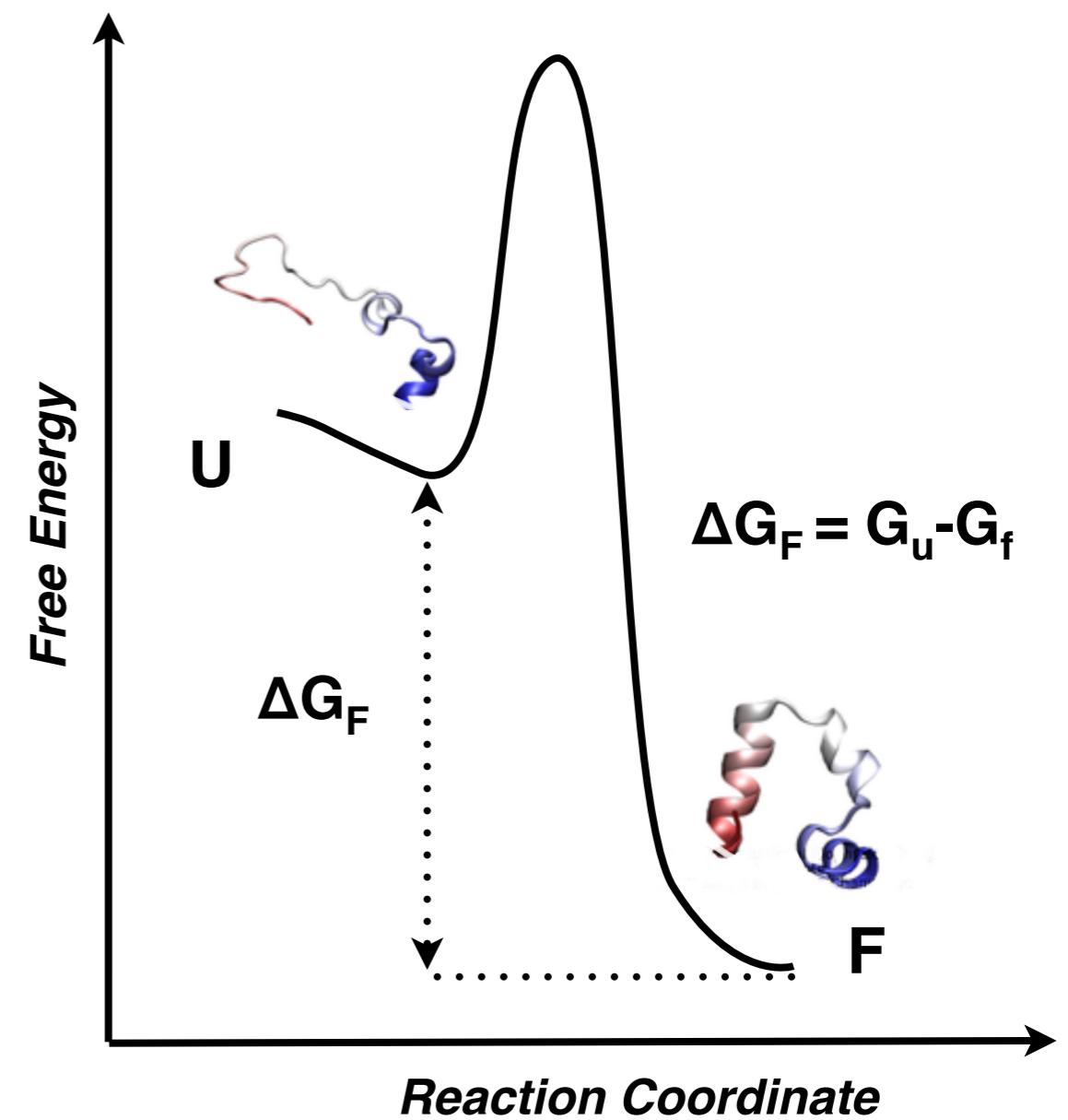
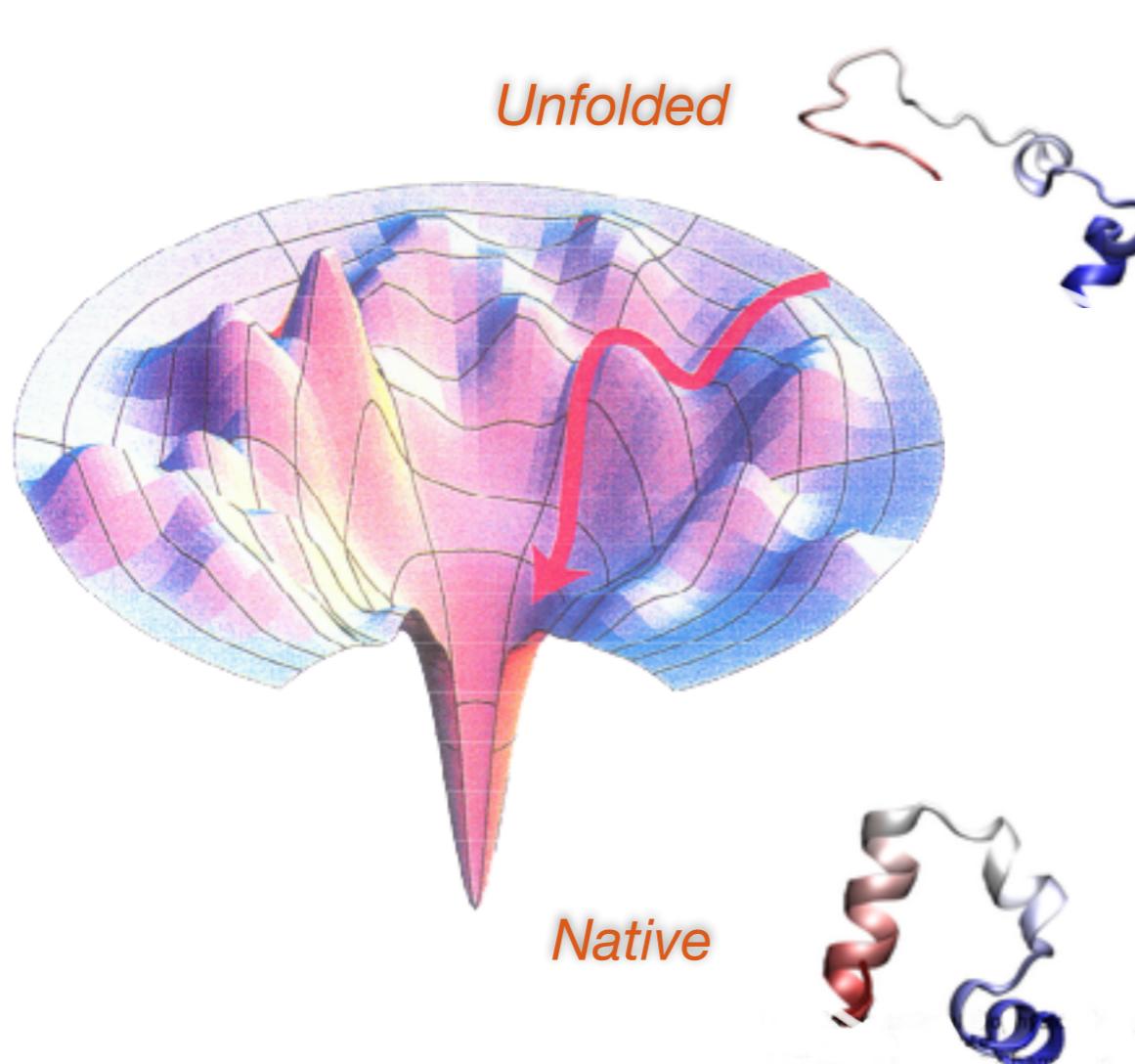
# Aspects of the same problem

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.

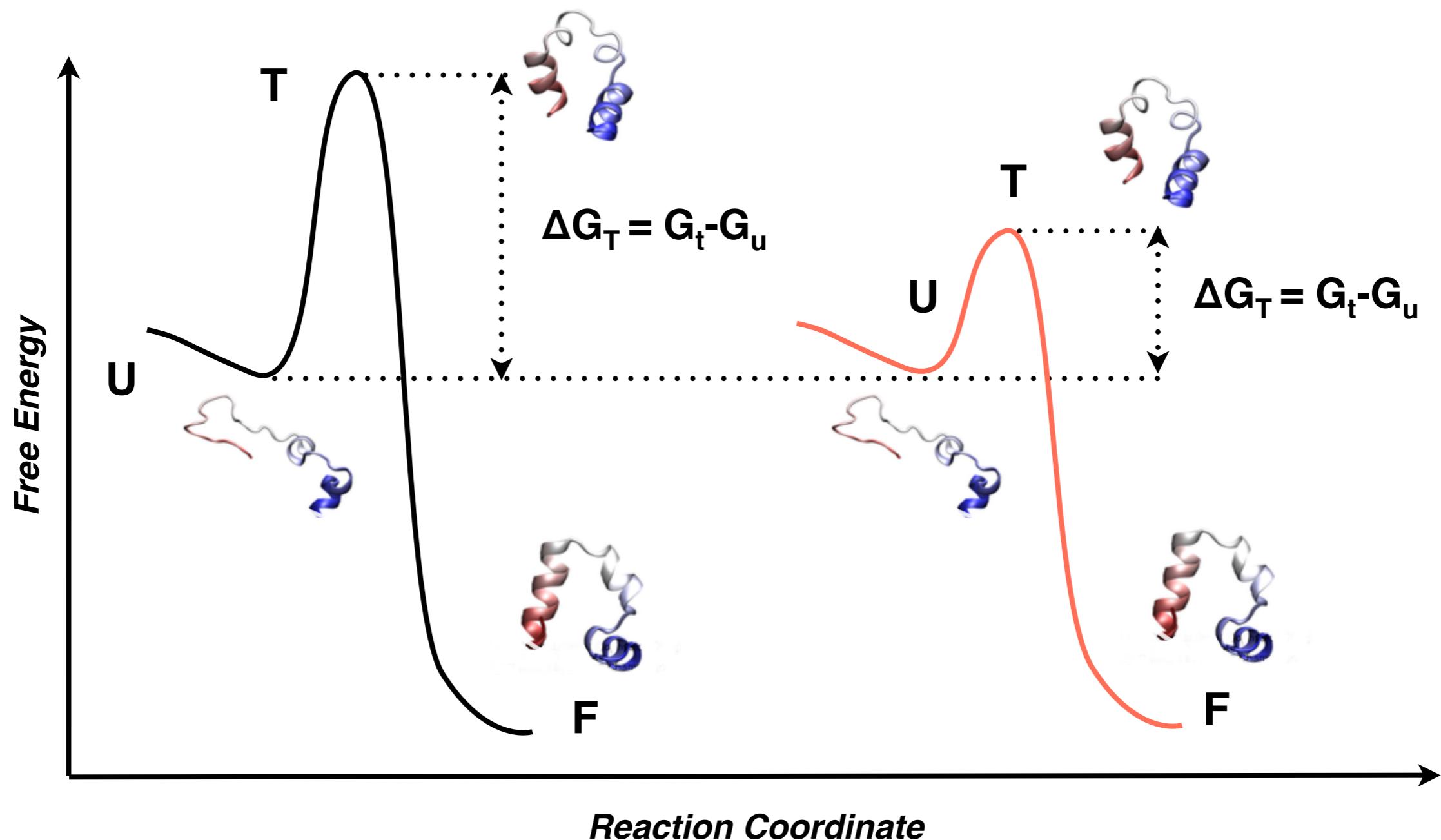
# Folding and stability

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).



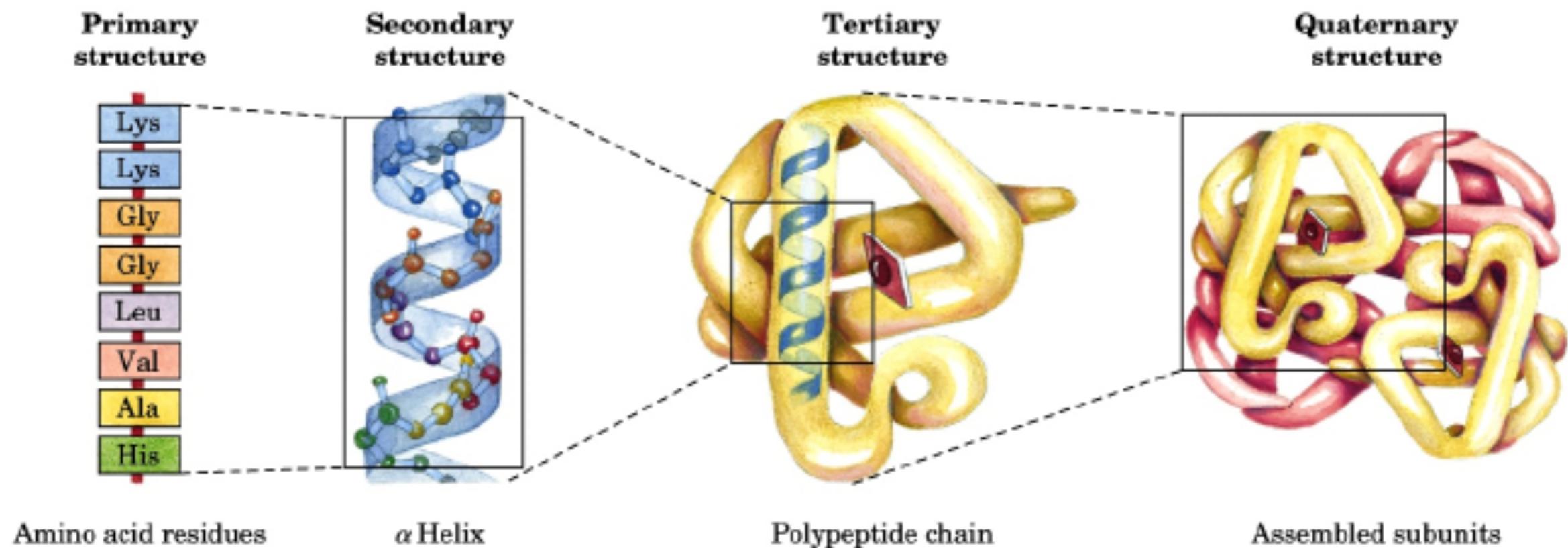
# Folding kinetics

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



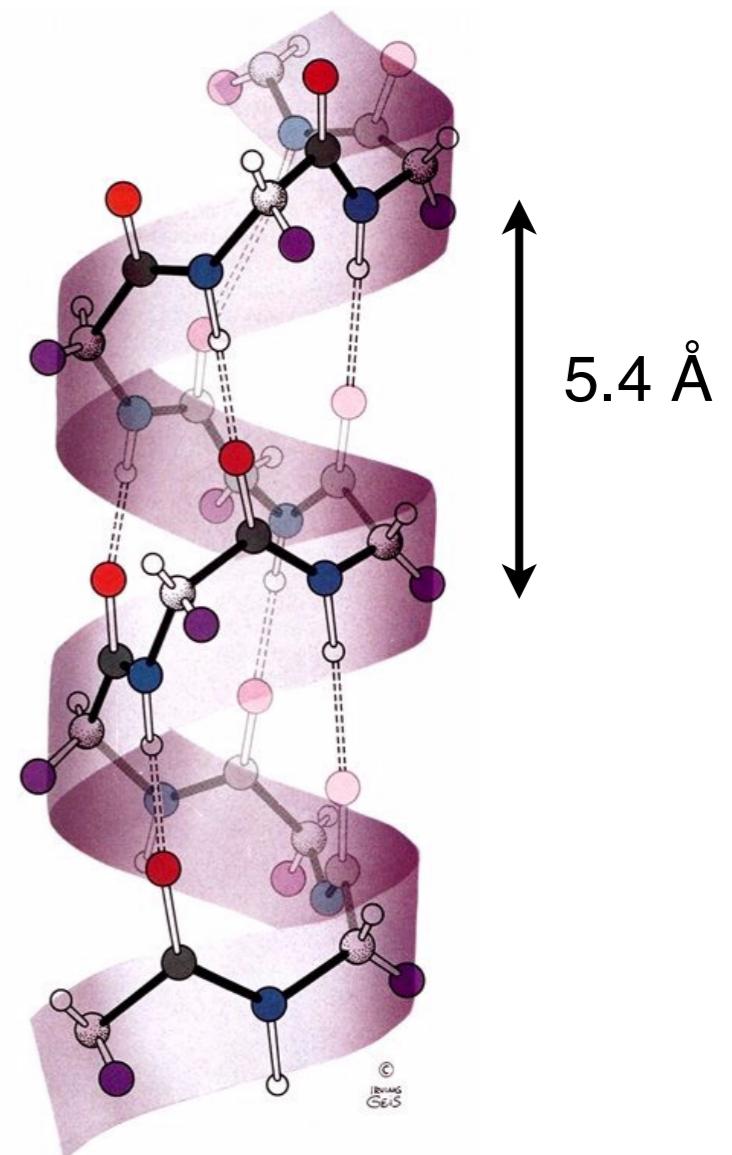
# Hierarchical organization of protein structure

Protein structure is defined by four levels of hierarchical organization.



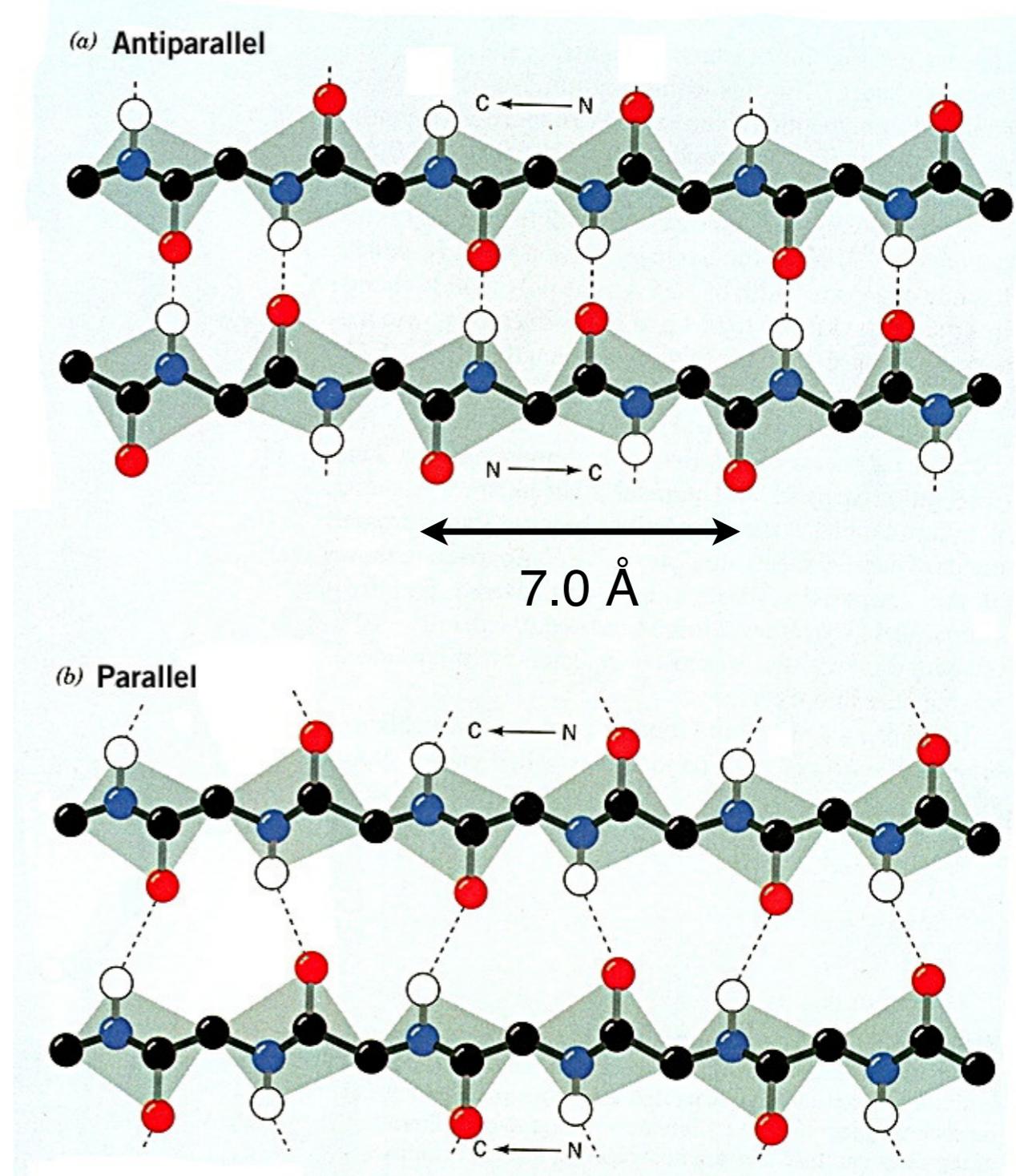
# Secondary structure (I)

- 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)

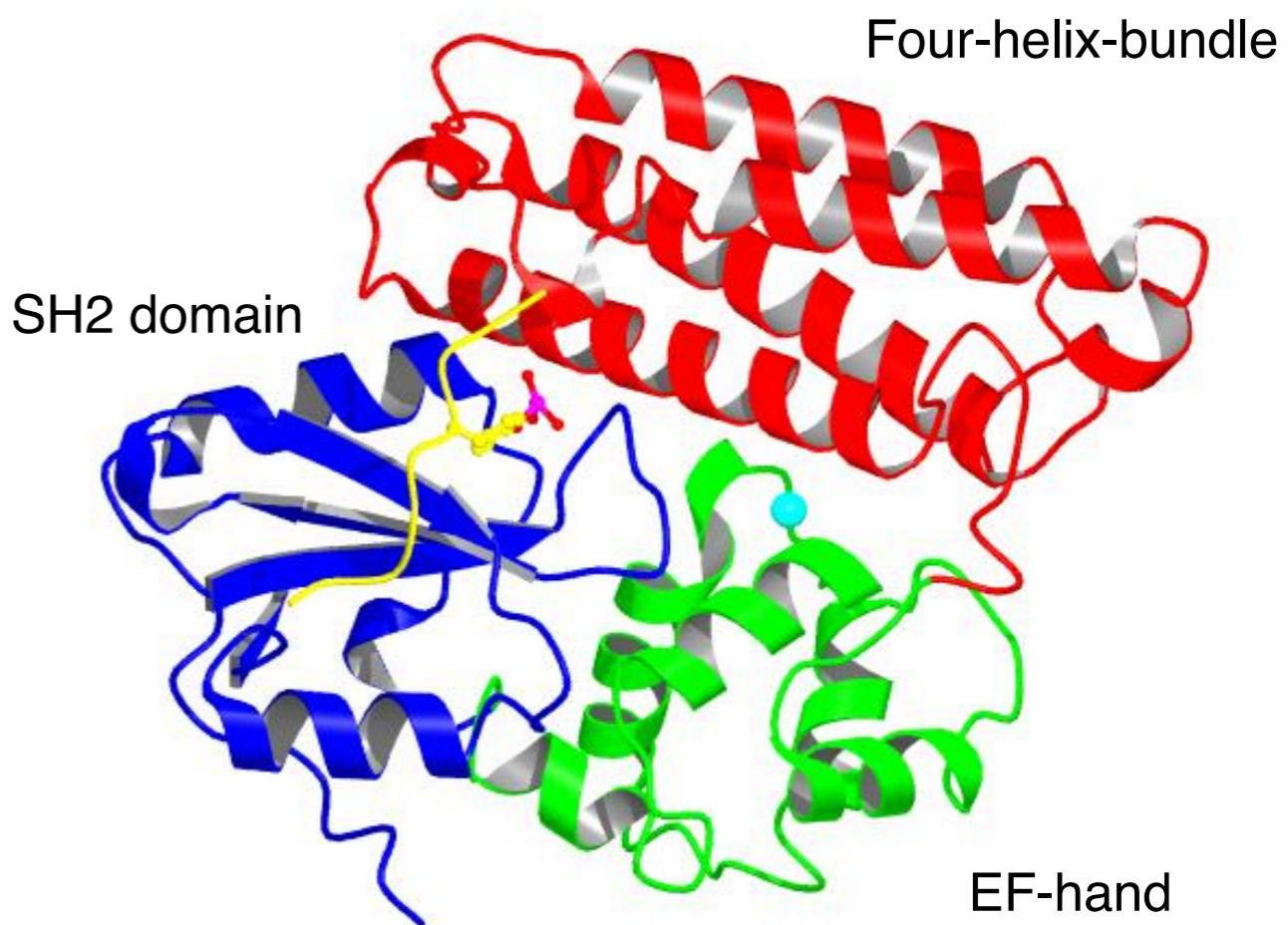
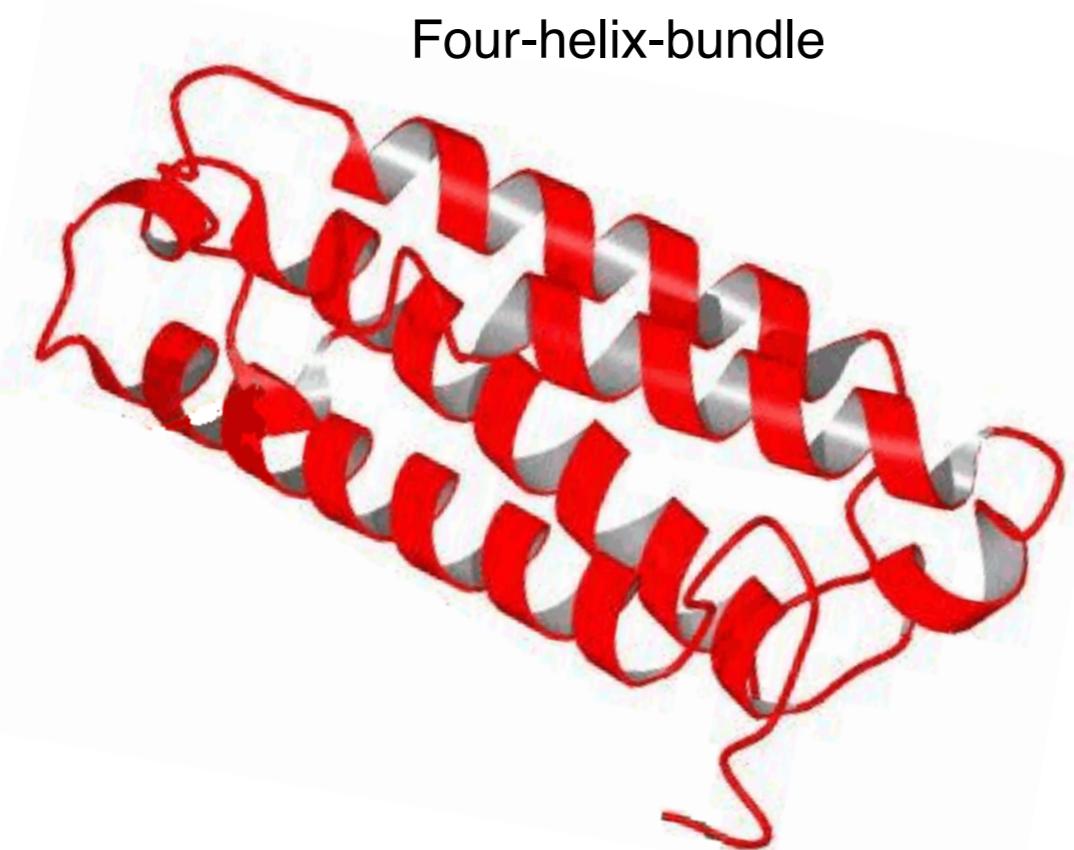
- Typical  $\phi$ ,  $\psi$  values for residues in  $\beta$ -sheet are around  $140^\circ$ ,  $-130^\circ$
- Side chains of neighboring residues project in opposite directions.
- The polypeptide is in a more extended conformation.
- Parallel  $\beta$ -sheets are less stable than anti-parallel  $\beta$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



# More complex structures

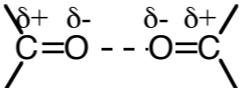
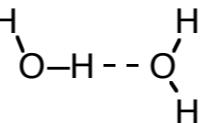
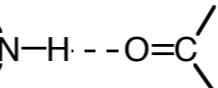
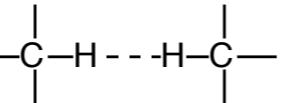
The arrangements of secondary structural elements form the Tertiary Structure of the protein.

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.



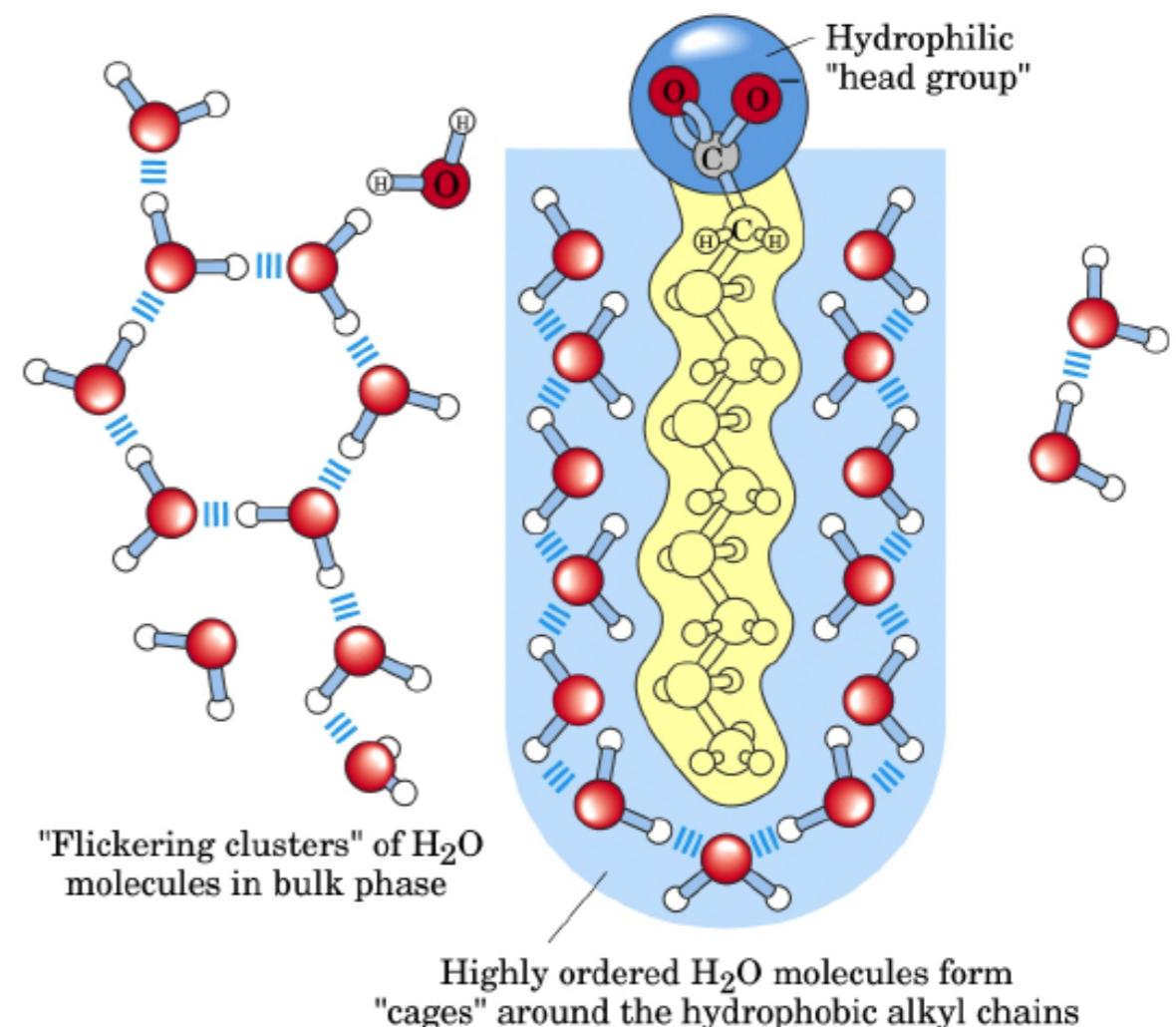
# Folding interactions

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 $\text{—COO} \cdots \text{N}^+\text{H}_3\text{—}$	-5	-1
	Dipole-dipole 	+0.3	
Hydrogen bond	Water 	-4	
	Protein backbone 	-3	
Dispersion forces	Aliphatic hydrogen 	-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.



# The Protein Data Bank

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

The screenshot shows the main interface of the RCSB PDB website. At the top, a dark blue header bar contains the "RCSB PDB" logo and navigation links for Deposit, Search, Visualize, Analyze, Download, Learn, and More, along with a "MyPDB Login" button. Below the header is a search bar with the placeholder "Search by PDB ID, author, macromolecule, sequence, or ligands" and a "Go" button. To the left of the search bar is the "PDB PROTEIN DATA BANK" logo and a brief description: "An Information Portal to 106517 Biological Macromolecular Structures". Below the search bar are links to "PDB-101", "EMDataBank", "NUCLEIC ACID DATABASE", and "StructuralBiology Knowledgebase". On the right side of the header, there are social media icons for Facebook, Twitter, YouTube, and mobile devices. The main content area features a large banner image of a protein structure. On the left, a sidebar menu lists "Welcome", "Deposit", "Search", "Visualize", "Analyze", "Download", and "Learn". The central column has a section titled "A Structural View of Biology" with text about the resource's purpose and history. It also includes a "Structure and Health Focus: Ebola Virus Proteins" section with a "Video Tour" and a "Molecule of the Month Article". On the right, a section titled "February Molecule of the Month" displays a 3D model of the Insulin Receptor.

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB Login

An Information Portal to 106517 Biological Macromolecular Structures

PDB-101 Worldwide Protein Data Bank EMDataBank Nucleic Acid Database StructuralBiology Knowledgebase

Search by PDB ID, author, macromolecule, sequence, or ligands Go

Advanced Search | Browse by Annotations

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.

Structure and Health Focus: Ebola Virus Proteins

Video Tour Molecule of the Month Article

February Molecule of the Month

Insulin Receptor

<http://www.pdb.org>

# CDK6-P16INK4A

Mechanism of CDK6 inhibition from the complex with tumor suppressor P16INK4A.

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB Login

**RCSB PDB** PROTEIN DATA BANK An Information Portal to 106710 Biological Macromolecular Structures

PDB-101 Worldwide Protein Data Bank EMDDataBank Nucleic Acid Database StructuralBiology Knowledgebase

Search by PDB ID, author, macromolecule, sequence, or ligands Go Advanced Search | Browse by Annotations

Summary 3D View Sequence Annotations Seq. Similarity 3D Similarity Literature Biol. & Chem. Methods Links

**MECHANISM OF G1 CYCLIN DEPENDENT KINASE INHIBITION FROM THE STRUCTURE OF THE CDK6-P16INK4A TUMOR SUPPRESSOR COMPLEX**

1BI7 Display Files Download Files Download Citation

DOI:10.2210/pdb1bi7/pdb

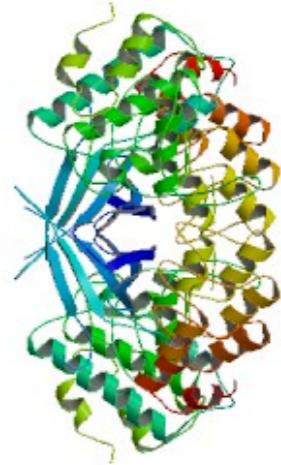
**Primary Citation**

Structural basis for inhibition of the cyclin-dependent kinase Cdk6 by the tumour suppressor p16INK4a.  
Russo, A.A., Tong, L., Lee, J.O., Jeffrey, P.D., Pavletich, N.P.  
Journal: (1998) Nature 395: 237-243  
PubMed: 9751050 DOI: 10.1038/26155 Search Related Articles in PubMed

**PubMed Abstract:**

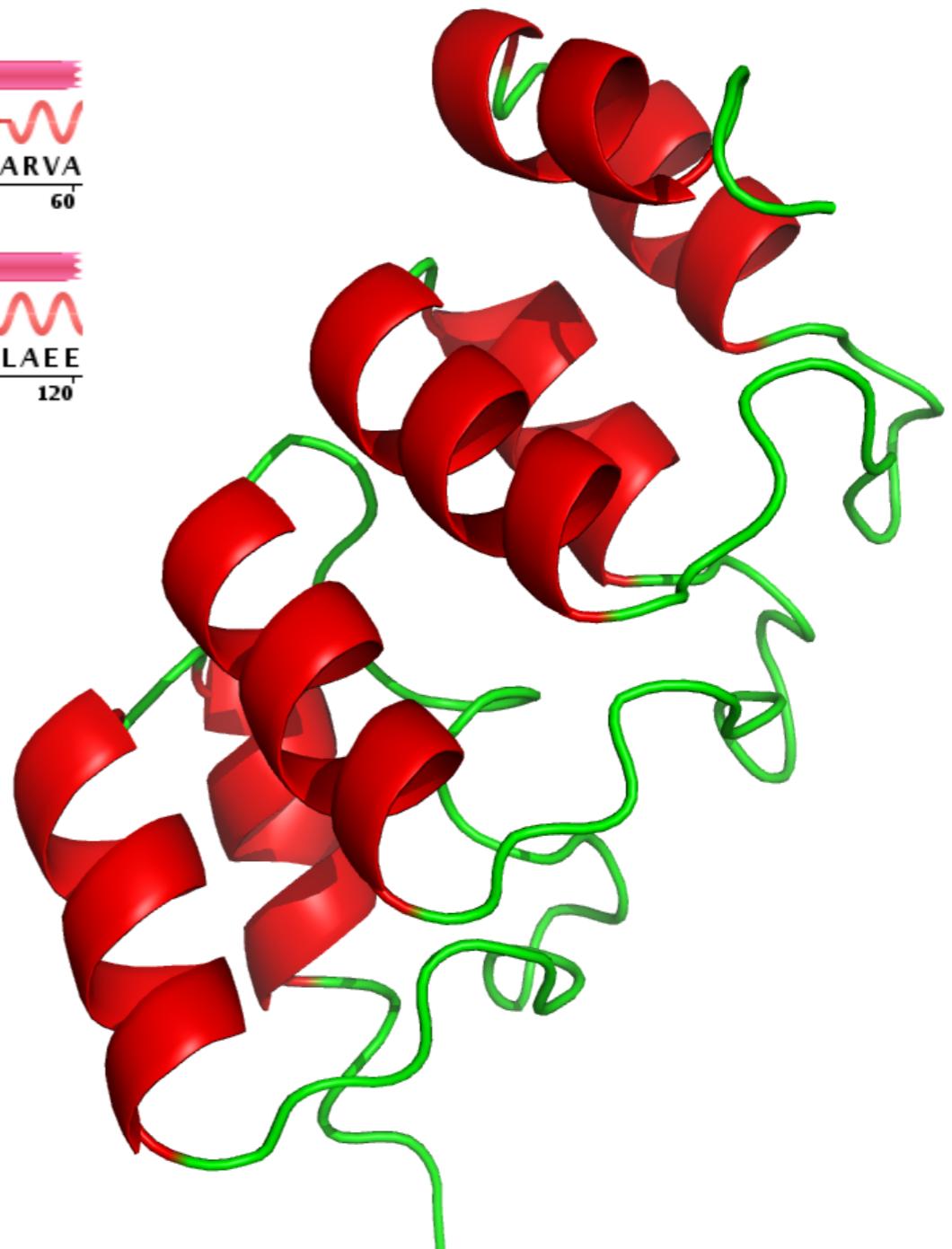
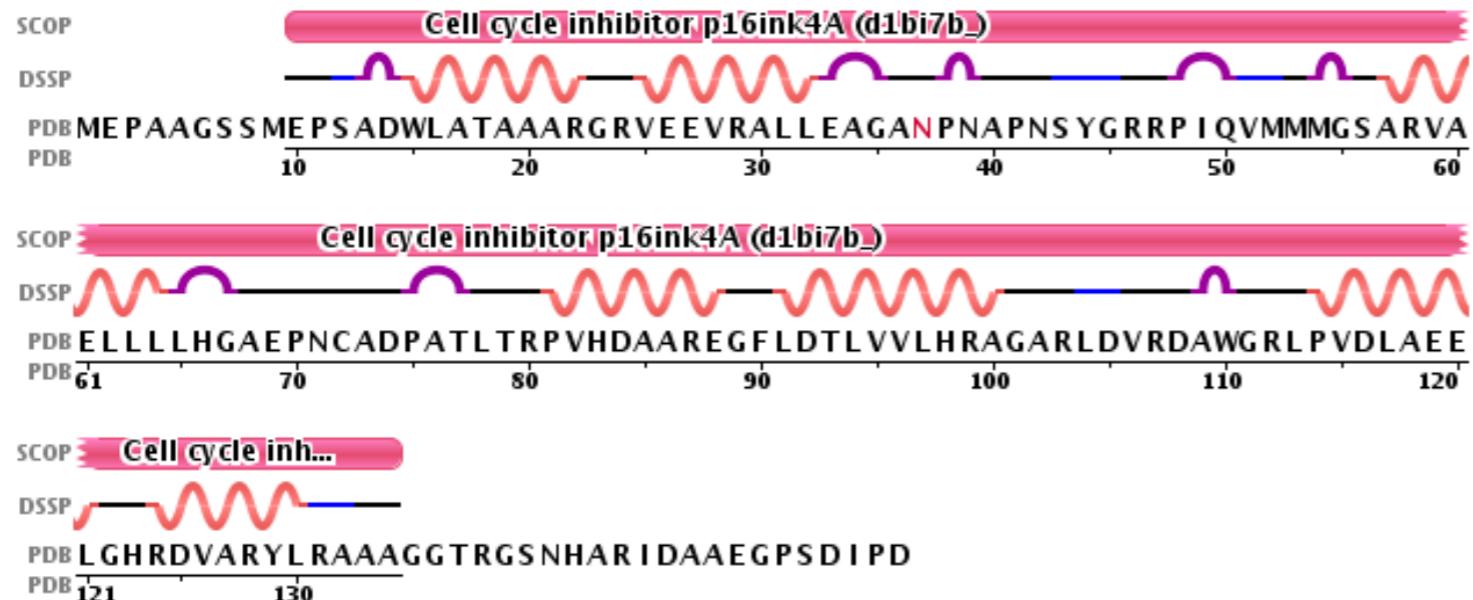
The cyclin-dependent kinases 4 and 6 (Cdk4/6) that control the G1 phase of the cell cycle and their inhibitor, the p16INK4a tumour suppressor, have a central role in cell proliferation and in tumorigenesis. The structures of Cdk6 bound to p16INK4a... [Read More & Search PubMed Abstracts]

**Biological Assembly**



# P16INK4A

The P16INK4A is a tumor suppressor protein with 7 helices.



# PDB data

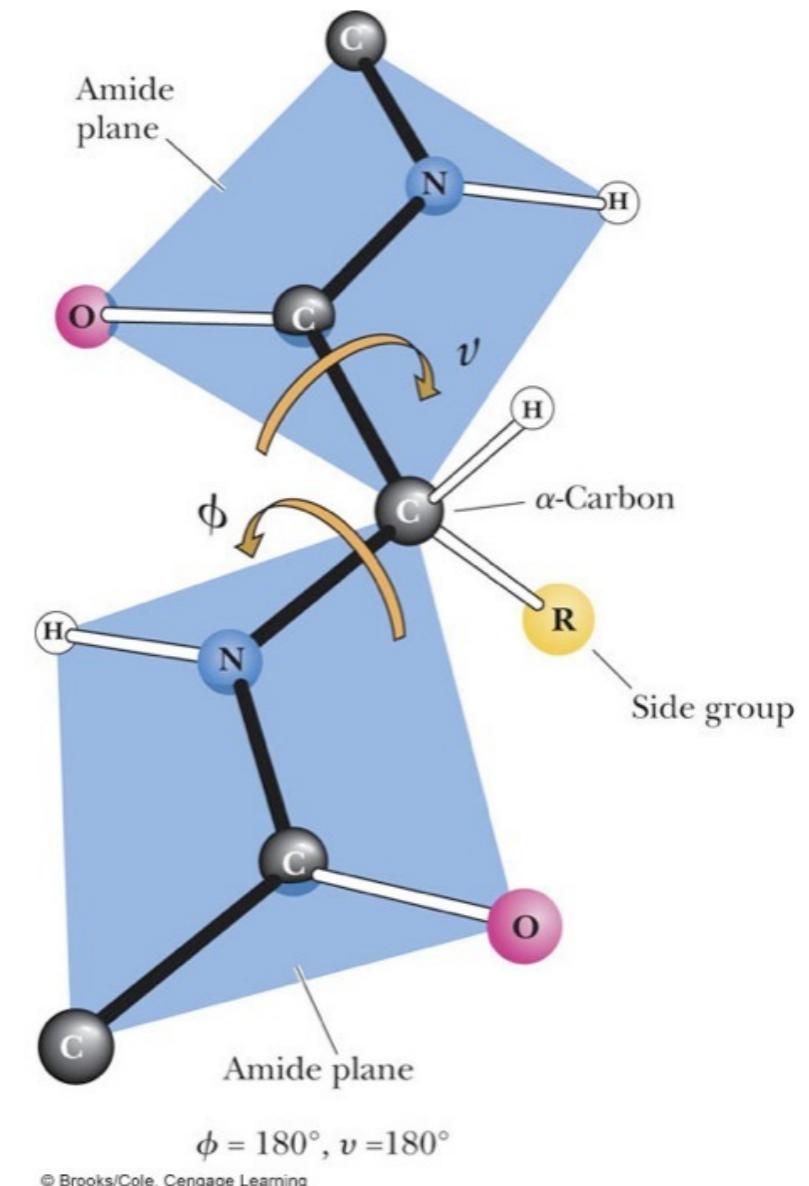
The most important information are the atomic coordinates.

	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

# Defining protein structure

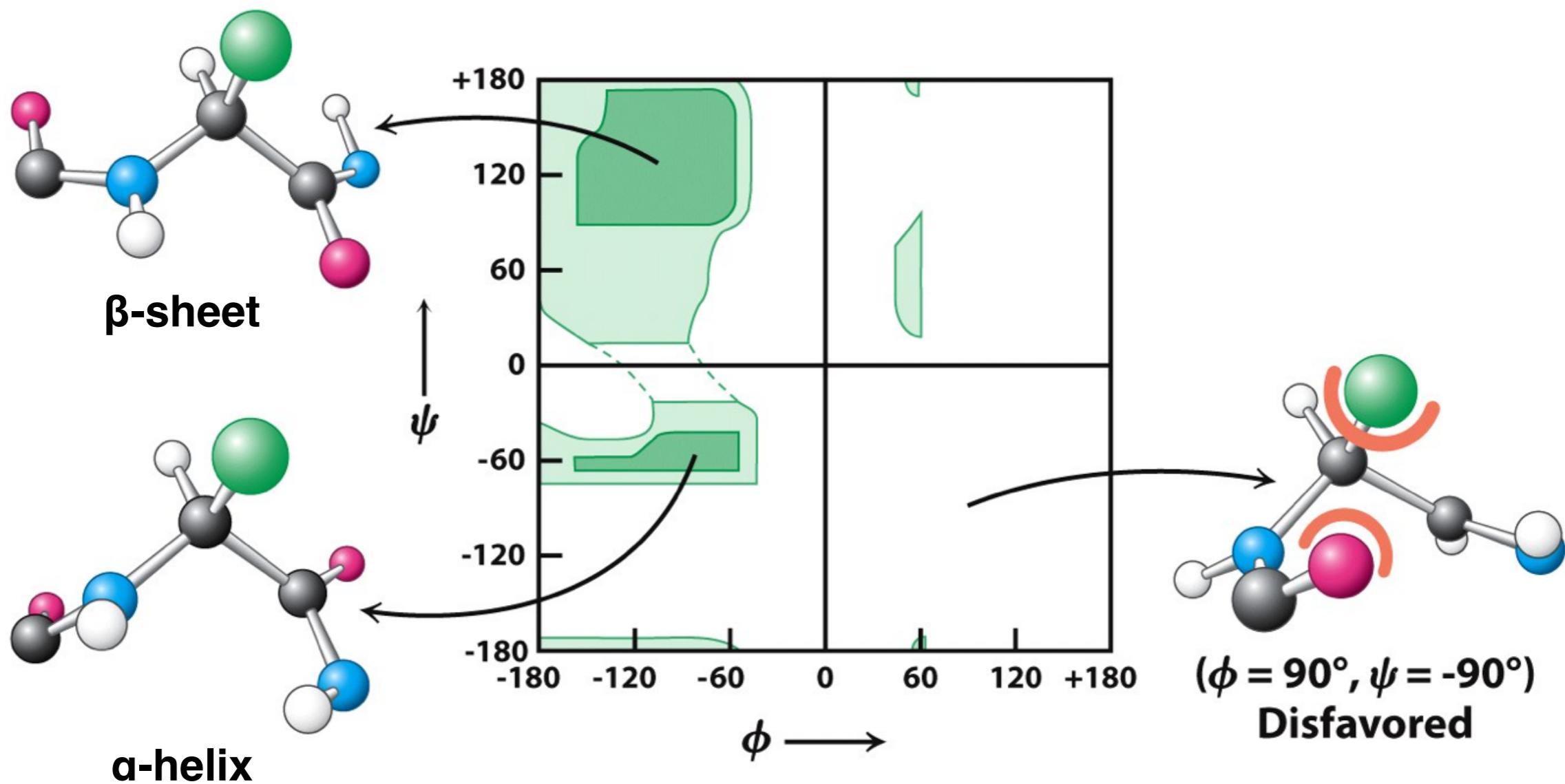
Basic information for the characterization of the protein three-dimensional structures are:

- $\phi, \psi$  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.



# 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**

DSSP: <ftp://ftp.cmbi.ru.nl/pub/software/dssp>  
more details at <http://www.cmbi.ru.nl/dssp.html>

# Problem 1a

Write a program that parse the DSSP file and for each residue extract:

- the secondary structure (col: 17)
- the solvent accessible area (cols: 36-38)
- phi and psi angles (cols: 104-109 and 110-115)

The program groups the different types of secondary structure in the three main ones (Helix, Beta and Coil) and calculate the relative solvent accessible area.

```
Norm_Acc={ "A" :106.0,   "B" :160.0,  
          "C" :135.0,   "D" :163.0,   "E" :194.0,  
          "F" :197.0,   "G" : 84.0,    "H" :184.0,  
          "I" :169.0,   "K" :205.0,   "L" :164.0,  
          "M" :188.0,   "N" :157.0,   "P" :136.0,  
          "Q" :198.0,   "R" :248.0,   "S" :130.0,  
          "T" :142.0,   "V" :142.0,   "W" :227.0,  
          "X" :180.0,   "Y" :222.0,   "Z" :196.0}
```

# Problem 1b

Write a script that takes in input a list of mutations and a DSSP file and chain, and returns for each mutation the secondary structure and the relative solvent accessible area.

How many mutated sites occurs in buried regions (relative solvent accessible area<20%)?

Run the script on the DSSPs obtained from the whole PDB and only from chain B to find possible mutation at the interface.