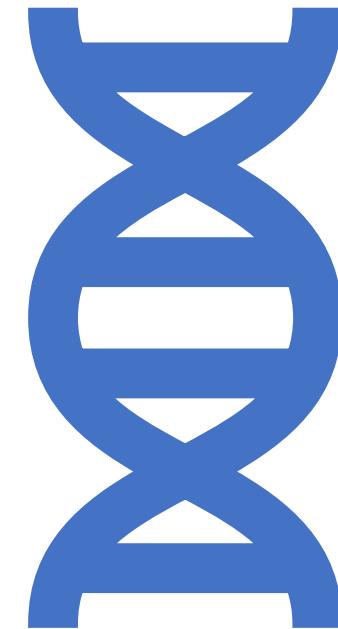


Structural Bioinformatics:

MSc Bioinformatics Cranfield

Dr Elliot Drew

e.drew@qmul.ac.uk



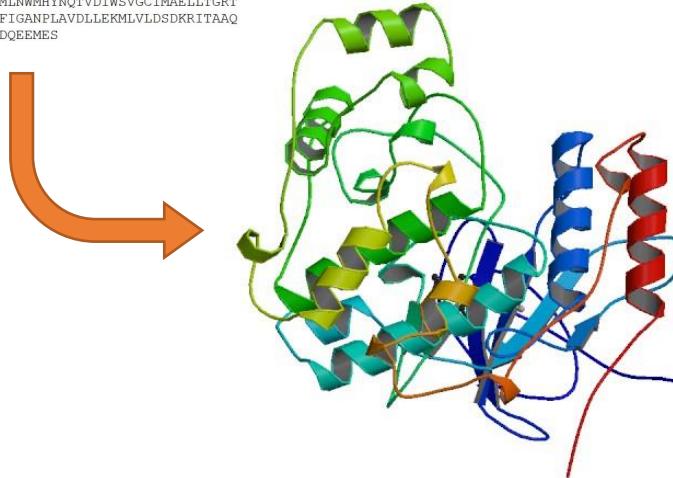
What is structural bioinformatics

It is the branch of bioinformatics related to the
analysis and prediction
of the 3D structures of biological macromolecules

What is structural bioinformatics

```
>1BL6:A|PDBID|CHAIN|SEQUENCE  
GSSHHHHHSSGLVPRGSMSQERPTFYRQEINKTIWEVPERYQNLSPVGSGAYGSVCAAFDTKTGLRAVKKLSRPFQS  
IHHAKRTYRELRLLKHMKHENVIQLLDVFTPARSLEEFNDVYLVTIIMGADLNIVKCQKLTDHVQFLIYQILRGLKYI  
HSADIHRDLKPSNLAVNEDCELKILDFGLARHTDEMTGYVATRWYAPEIMLNWMHYNQTVDIWSVGCIMAELLTGRT  
LFPGTGDHIDQQLKLILRLVGTGPGAEILLKKISSESARNYIQSLTQMPKMFANVFIGANFLAVDLEKMLVLSDDKRITAAQ  
ALAHAYFAQYHPDDEPVADPYDQSFESESRDLIDEWKSITYDEVISFVPPPLDQEEMES
```

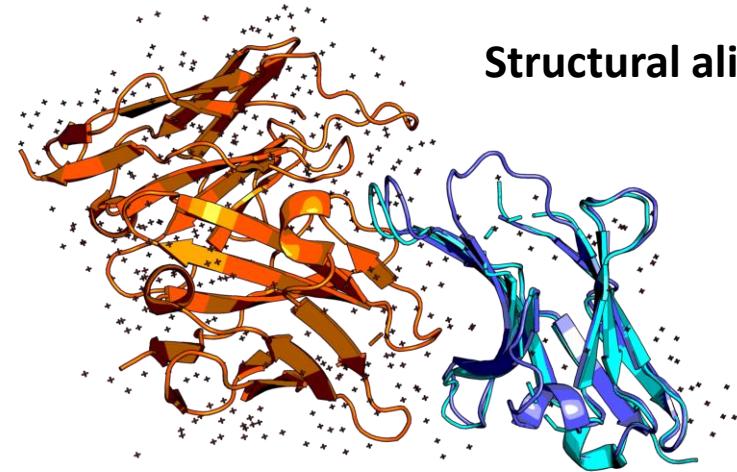
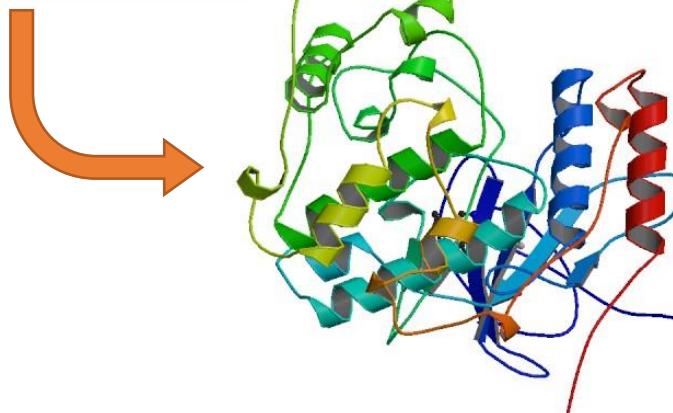
Predict Structure



What is structural bioinformatics

```
>1BL6:A|PDBID|CHAIN|SEQUENCE  
GSSHHHHHSSGLVPRGSMSQERPTFYRQEINKTIWEVPERYQNLSPVGSGAYGSVCAAFDTKTGLRAVKKLSRPFQS  
IIHAKRTYRELRLLKHMKHENVIGLLDVFTPARSLEEFNDVYLVTIIMGADLNIVKCQKLTDHVQFLIYQILRGLKYI  
HSADIHRDLKPSNLAVNEDCELKILDFGLARHTDDEMTGYVATRWYAPEIMLNWMHYNQTVDIWSVGCIMAELLTGRT  
LFPGTGDHIDQQLKLILRLVGTGPGAEILLKKISSESARNVIQSLTQMPKMFANVFIGANLAVDLEKMLVLSDKRITAAQ  
ALAHAYFAQYHDPDDEPVADPYDQSFESESDLIDEWKSLTYDEVISFVPPPLDQEEMES
```

Predict Structure

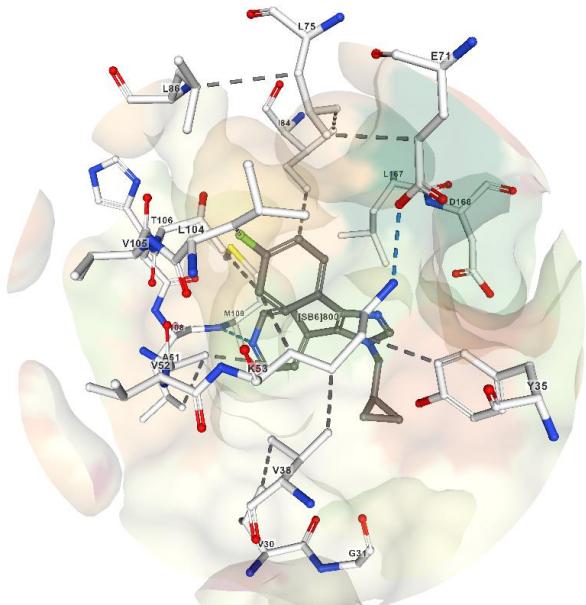


Structural alignment

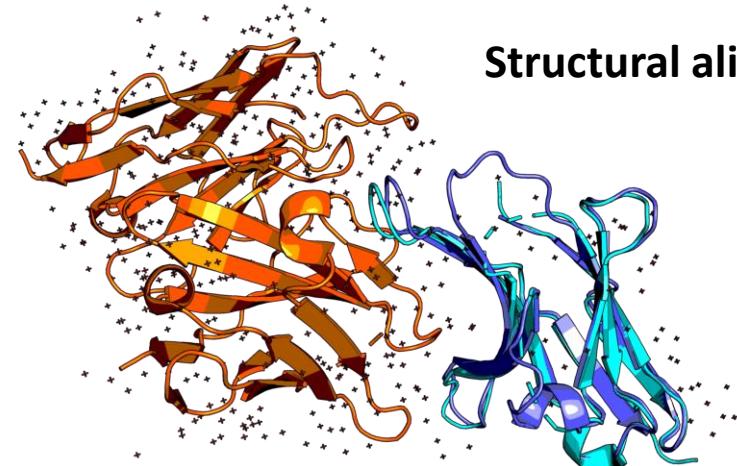
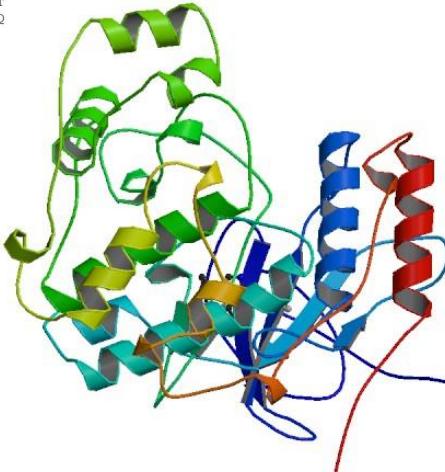
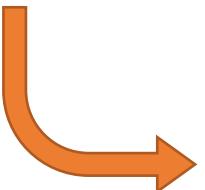
What is structural bioinformatics

```
>1BL6:A|PDBID|CHAIN|SEQUENCE  
GSSHHHHHSSGLVPRGSQMSQERTFYRQEINKTIWEVPERYQNLSPVGSGAYGSVCAAFDTKTLRVAVKKLSRPFQS  
IHHAKRTYRELRLLKHMKHENVIQLLDVFTPARSLEEFNDVYLVTIIMGADLNIVKCQKLTDHVQFLIYQILRGLKYI  
HSADIHRDLKPSNLAVNEDCELKILDGLARHTDEMTGYVATRWYAPEIMLNWMHYNQTVDIWSVGCIMAEELLGRT  
LFPGTGDHIDQQLKLILRLVGTGPGAEILLKKISSESARNVIQSLTQMPKMFANFIGANFLAVDLEKMLVLDSDKRITAAQ  
ALAHAYFAQYHPDDEPVADPYDQSFESESDLILDEWKSITYDEVISFVPPPLDQEEMES
```

Predict Structure



Predict Interactions

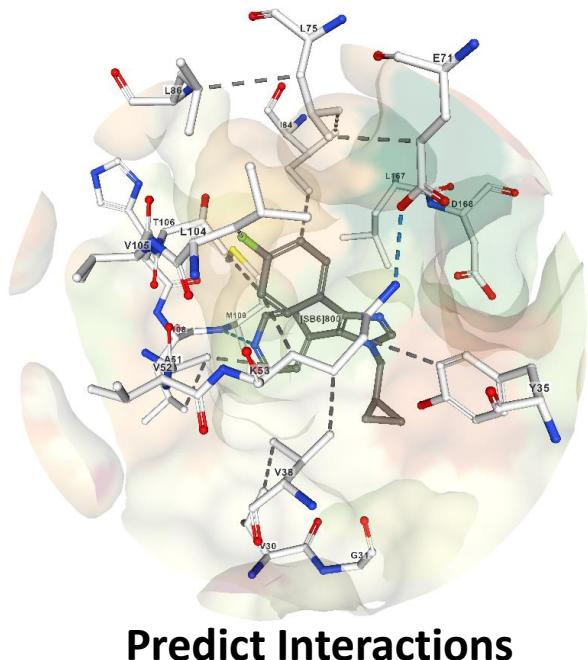


Structural alignment

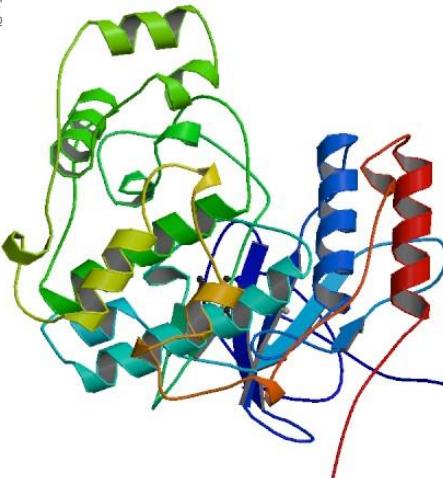
What is structural bioinformatics

>1BL6:A | PDBID | CHAIN | SEQUENCE
GSSHHHHHSSGLVPRGSHMSQERPTFYRQEINKTIWEVPERYQNLSPVGSGAYGSVCAAFDTKTLRVAVKKLSRPFQS
I1HAKRTYRELRLLKHMKHENVIQLLDVFTPARSLEEFNDVYLVTIIMGADLNIVKCQKLTDHVQFLIYQILRGLKYI
HSADI1HRLDKPSNLAVNEDCEKL1LDFGLARHTDEMTGYVATRWYAPEIMLNWMHYNQTVDIWSVGCIMAEELLGRT
LFPGTIDHIQDLKL1LRLVGTGPAEELLKKISSESARNVIQSLTQMPKMFANFVIGANLAVDLEKMLVLDSDKRITAAQ
ALAHAYFAQYHPDDEPVADPYDQSFESESDLIDEWKS1TYDEVISFVPPPLDQEEMES

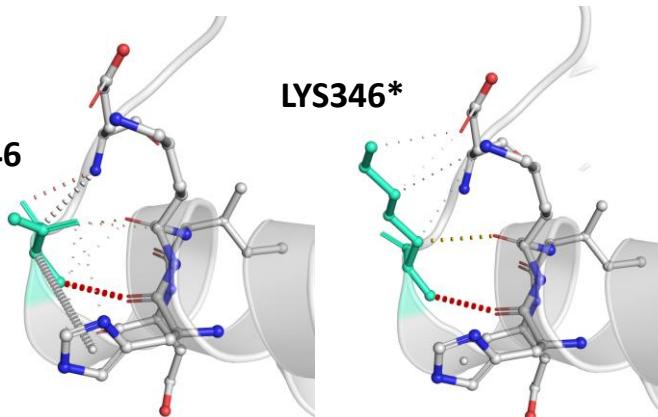
Predict Structure



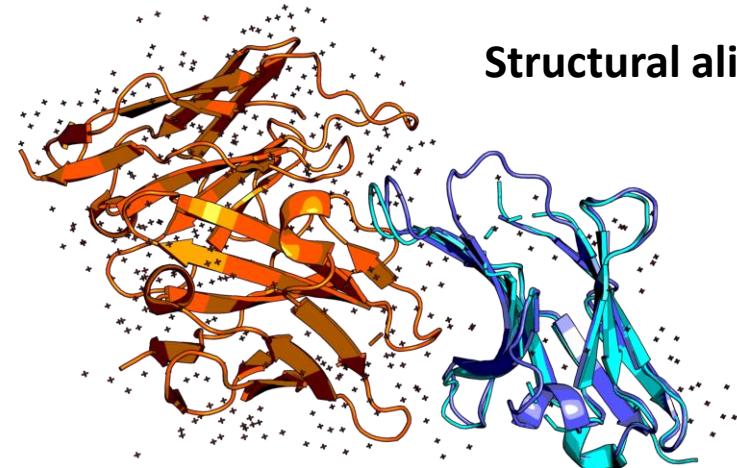
Predict Interactions



GLU346



Computational Mutagenesis

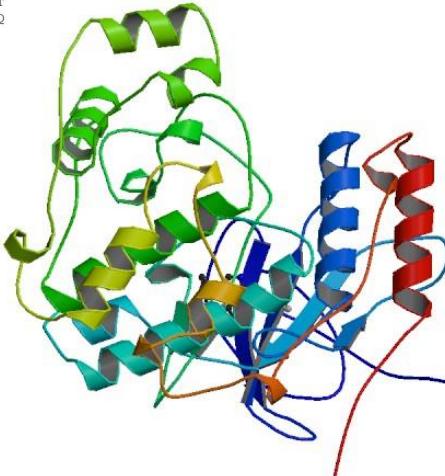
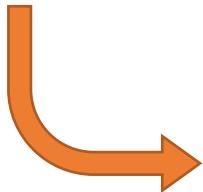
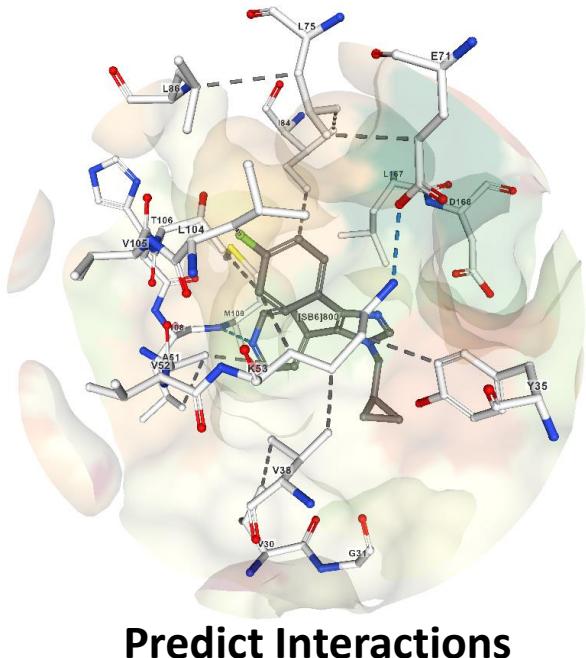


Structural alignment

What is structural bioinformatics

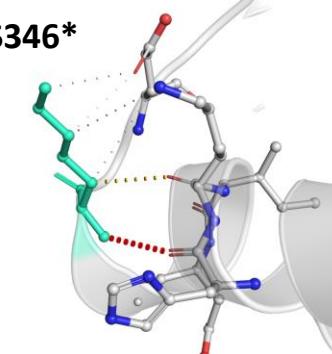
>1BL6:A | PDBID | CHAIN | SEQUENCE
GSSHHHHHSSGLVPRGSHMSQERTFYRQEINKTIEWVPERYQNLSPVGSGAYGSVCAAFDTKTLRVAVKKLSRPFQS
IHHAKRTYRELRLLKHMKHENVIQLLDVFTPARSLEEFNDVYLVTIIMGADLNIVKCQKLTDHVQFLIYQILRGLKYI
HSADIHRDLKPSNLAVNEDCEKLILDFGLARHTDEMTGYVATRWYAPEIMLNWMHYNQTVDIWSVGCIMAEELLGRT
LFPGTIDHIQDQLKLILRLVGTGPGAEILLKKISSESARNVIQSLTQMPKMFANFVIGANLAVDLEKMLVLDSDKRITAAQ
ALAHAYFAQYHPDDEPVADPYDQSFESESDLIDEWKSITYDEVISFVPPPLDQEEMES

Predict Structure

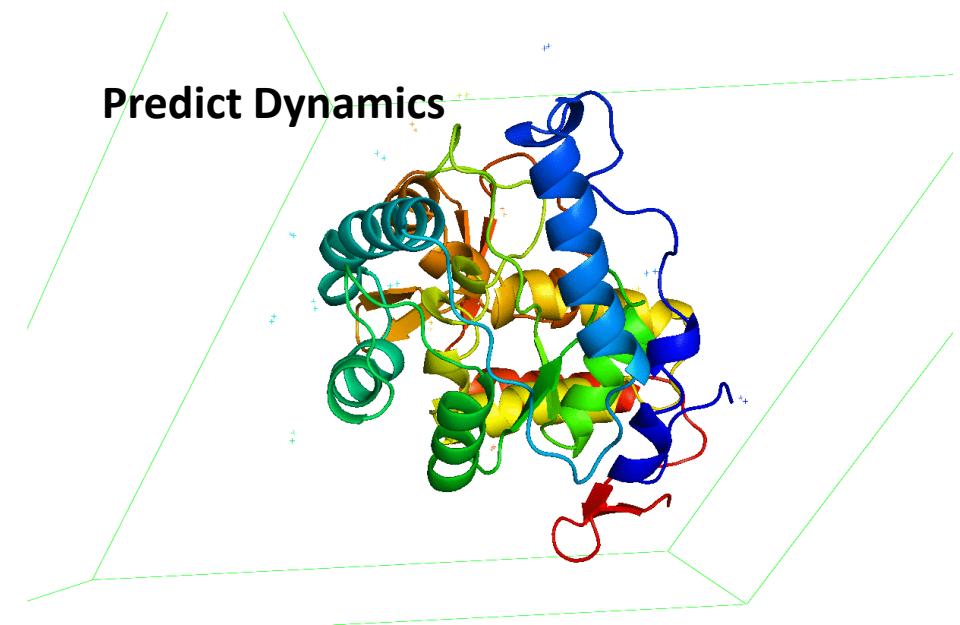


Computational Mutagenesis

GLU346



Predict Dynamics

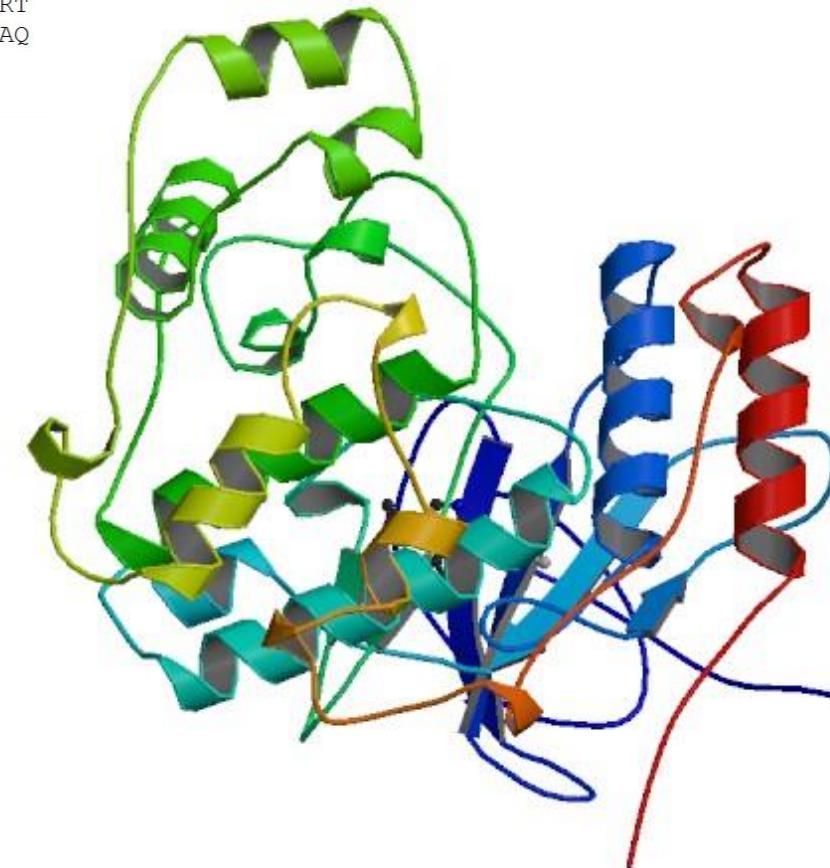
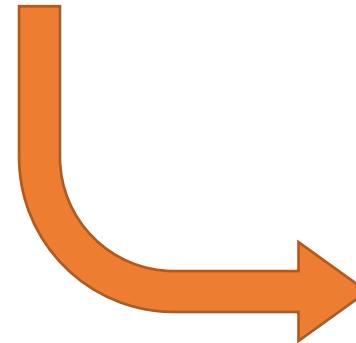


Structural alignment

What is structural bioinformatics

>1BL6:A|PDBID|CHAIN|SEQUENCE
GSSHHHHHSSGLVPRGHMSQERPTFYRQEINKTIWEVPERYQNLSVGSGAYGSVCAAFDTKGLRVAVKKLSRPFQS
IIHAKRTYRELRLLKHMKHENVIGLLDVFTPARSLEEFNDVYLVTLMGADLNNIVKCQKLDDHVQFLIYQILRGLKYI
HSADIIRDLKPSNLAVNEDCELKILDGLARHTDEMTGYVATRWYRAPEIMLNWMHYNQTVDIWSVGCIMAELLTGRT
LFPGTDHIDQLKLILRLVGTPGAELLKKISSESARNYIQSLTQMPKMNFANVFIGANPLAVDLLEKMLVLDSDKRITAAQ
ALAHAYFAQYHDPDDEPVADPYDQSFESRDLLIDEWKSLTYDEVISFVPPPLDQEEMES

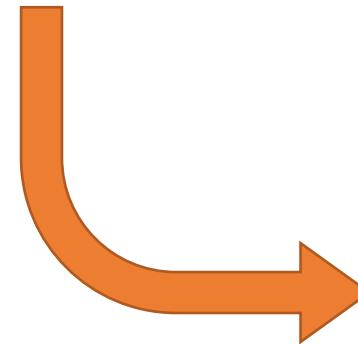
Predict Structure



What is structural bioinformatics

```
>1BL6:A|PDBID|CHAIN|SEQUENCE
GSSHHHHHSSGLVPRGHMSQERPTFYRQEINKTIWEVPERYQNLSPGSGAYGSVCAAFDTKGLRVAVKKLSPFQS
I IHAKRTYRELRLLKHMKHENVIGLLDVFTPARSLEEFNDVYLVTLMGADLNNIVKCQKLTDDHVQFLIYQILRGLKYI
HSADI IHRDLKPSNLAVNEDCELKILDGLARHTDEMTGYVATRWYRAPEIMLNWMHYNQTVDIWSVGCIMAELLTGRT
LFPGTDHIDQLKLILRLVGT PGAELLKKISSESARNYIQSLTQMPKMN FANVFIGANPLAVDLLEKMLVLDSDKRITAAQ
ALAHAYFAQYHDPDDEPVADPYDQSFESRDLLIDEWKSLTYDEVISFVPPPLDQEEMES
```

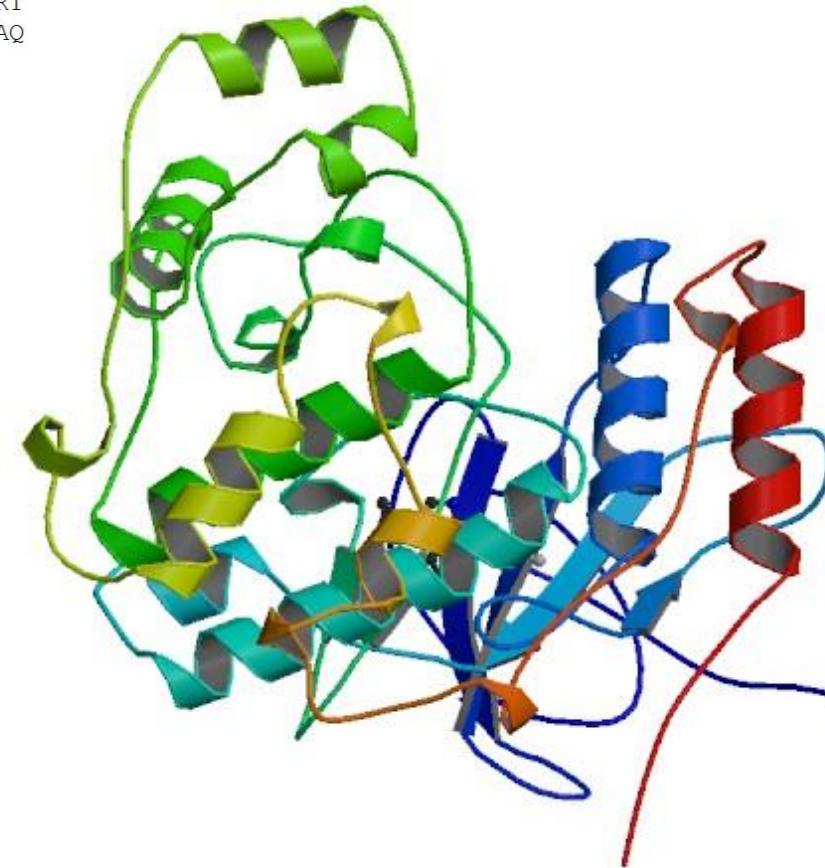
Predict Structure



195 million sequences in TrEMBL

vs

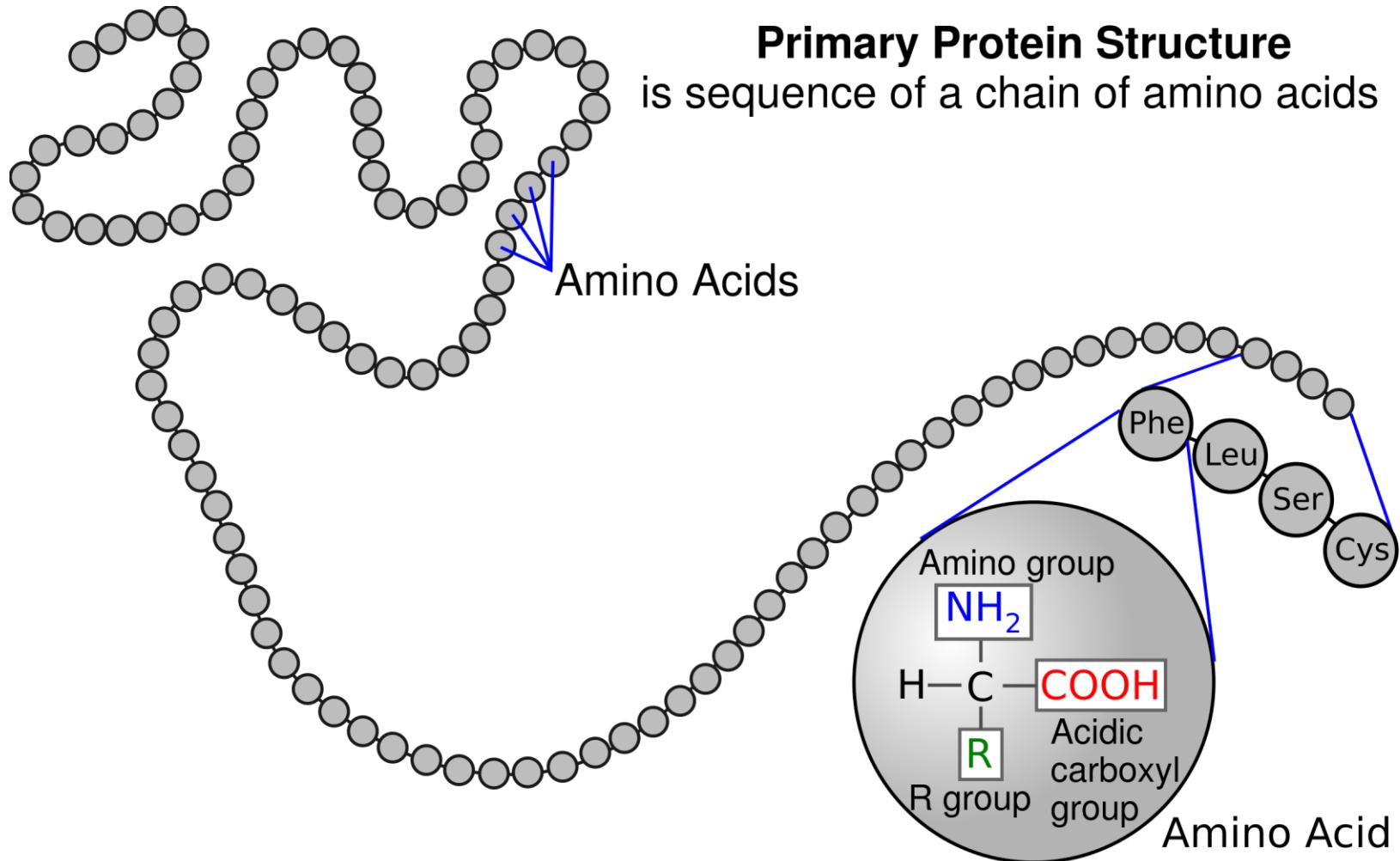
170,000 structures in Protein Data Bank



Protein Structure

Four levels of Protein structure:

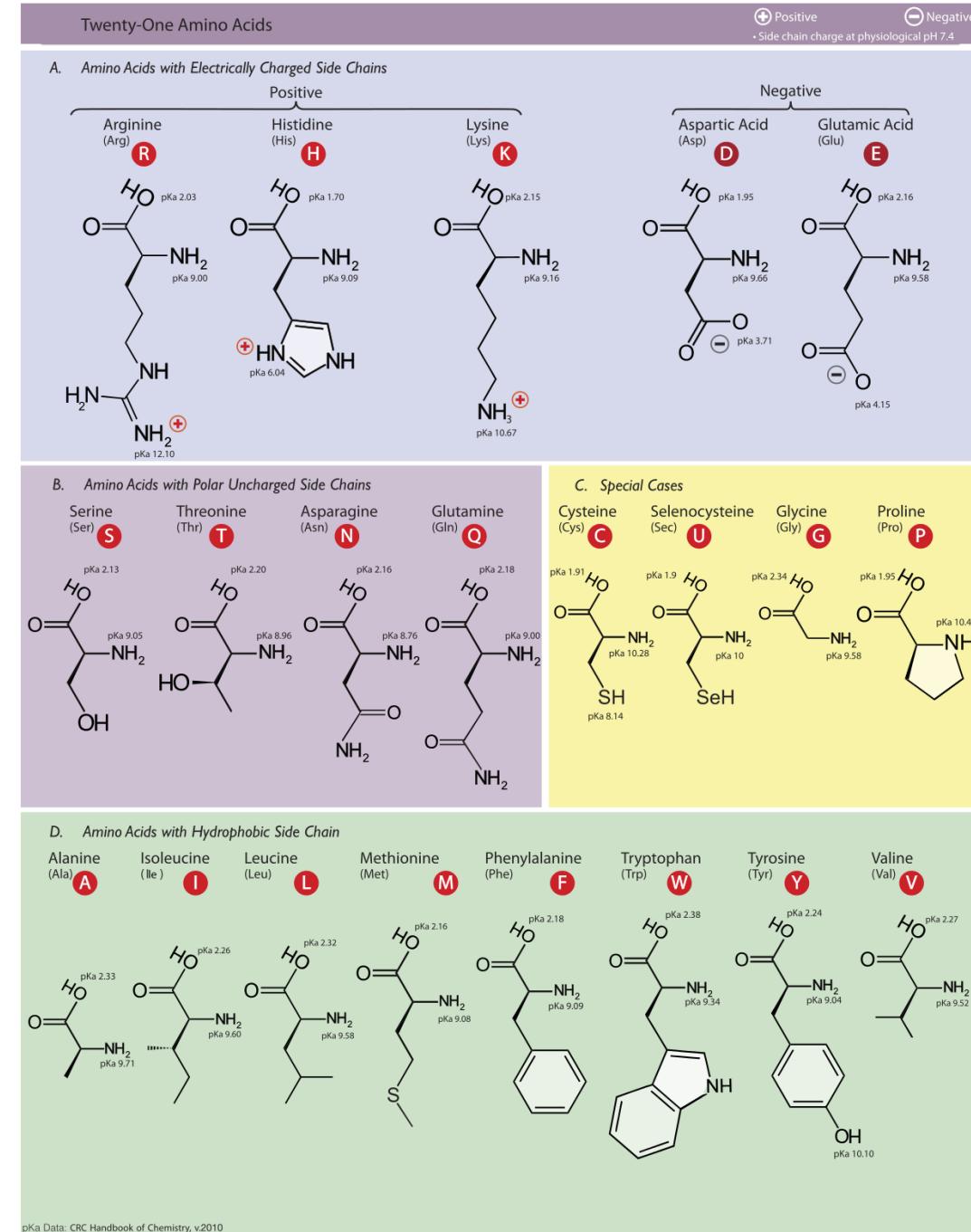
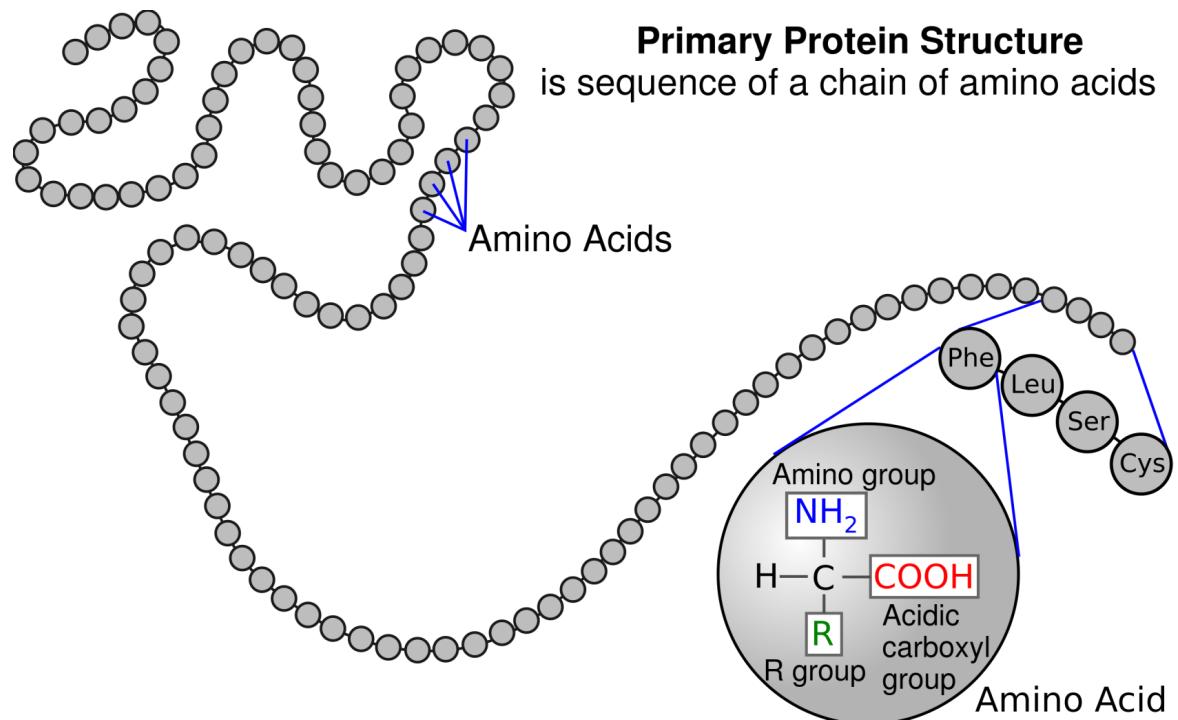
1. Primary



Protein Structure

Four levels of Protein structure:

1. Primary

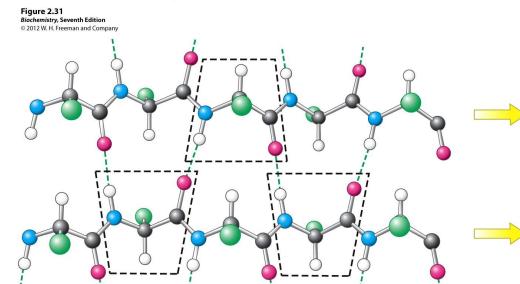
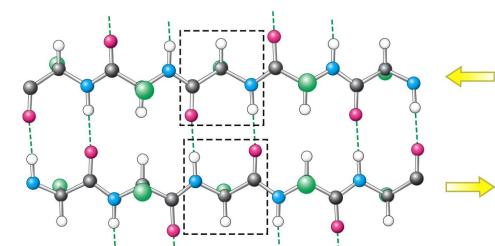
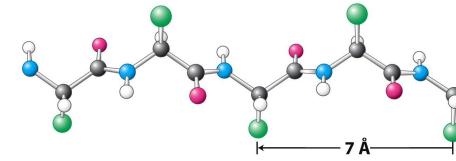
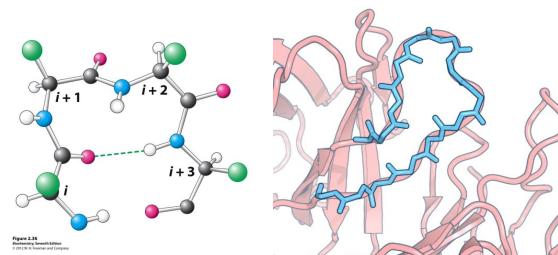
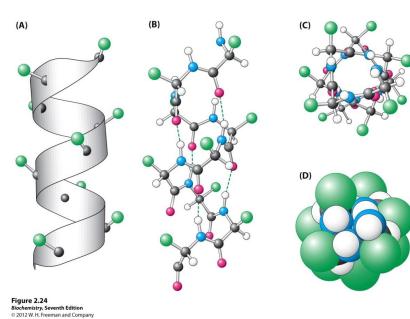


Protein Structure

Four levels of Protein structure:

1. Primary
2. Secondary

- 3D arrangement of local segments
- Hydrogen Bonds
- Phi/Psi angles
- “3 State” – H E O



https://www.youtube.com/watch?v=1usemtIYe_s

Protein Structure

Four levels of Protein structure:

1. Primary
2. Secondary

- 3D arrangement of local segments
- Hydrogen Bonds
- Phi/Psi angles
- “3 State” – H E O

Alpha Helix

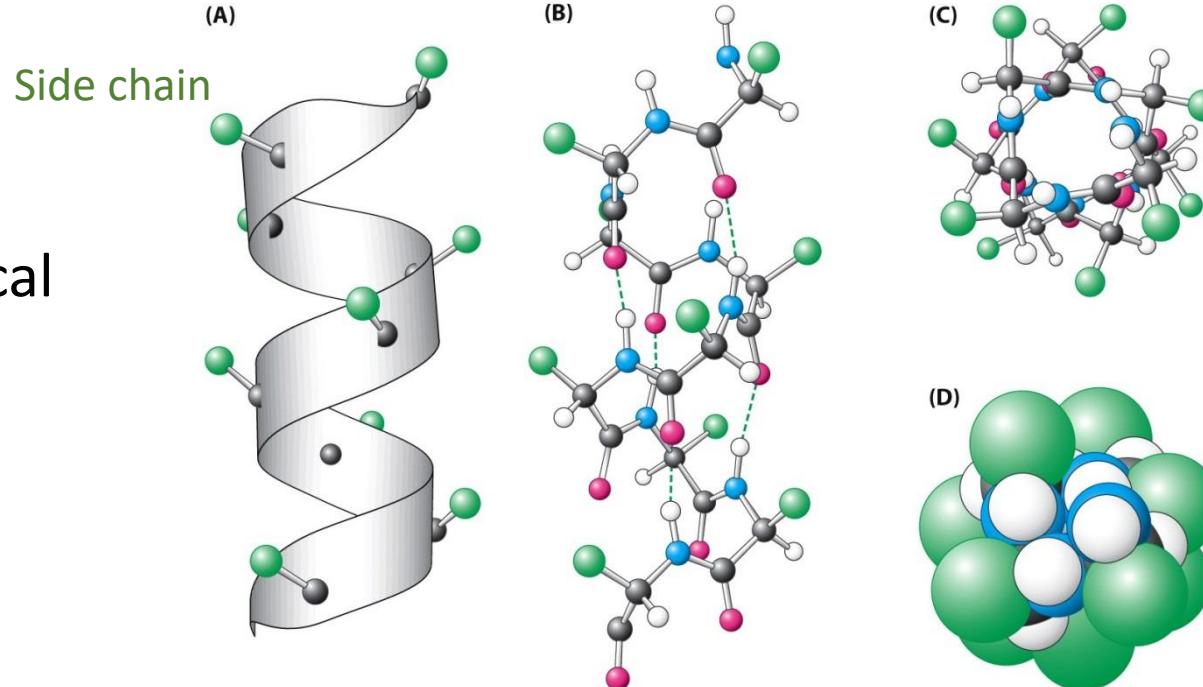


Figure 2.24
Biochemistry, Seventh Edition
© 2012 W. H. Freeman and Company

Protein Structure

Four levels of Protein structure:

1. Primary
2. Secondary
 - 3D arrangement of local segments
 - Hydrogen Bonds
 - Phi/Psi angles
 - “3 State” – H E O

Beta Strand

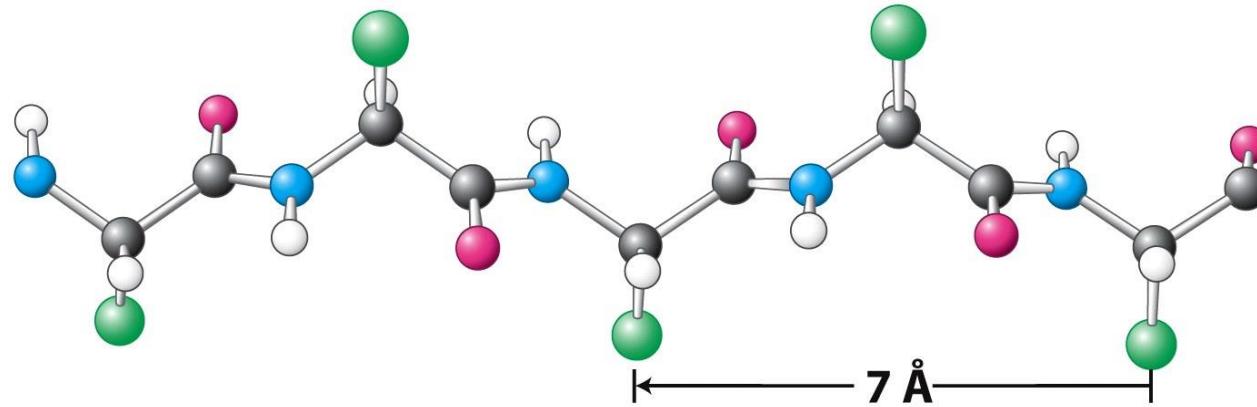


Figure 2.30
Biochemistry, Seventh Edition
© 2012 W. H. Freeman and Company

Protein Structure

Four levels of Protein structure:

1. Primary
2. Secondary
 - 3D arrangement of local segments
 - Hydrogen Bonds
 - Phi/Psi angles
 - “3 State” – H E O

Anti-Parallel Beta Sheet

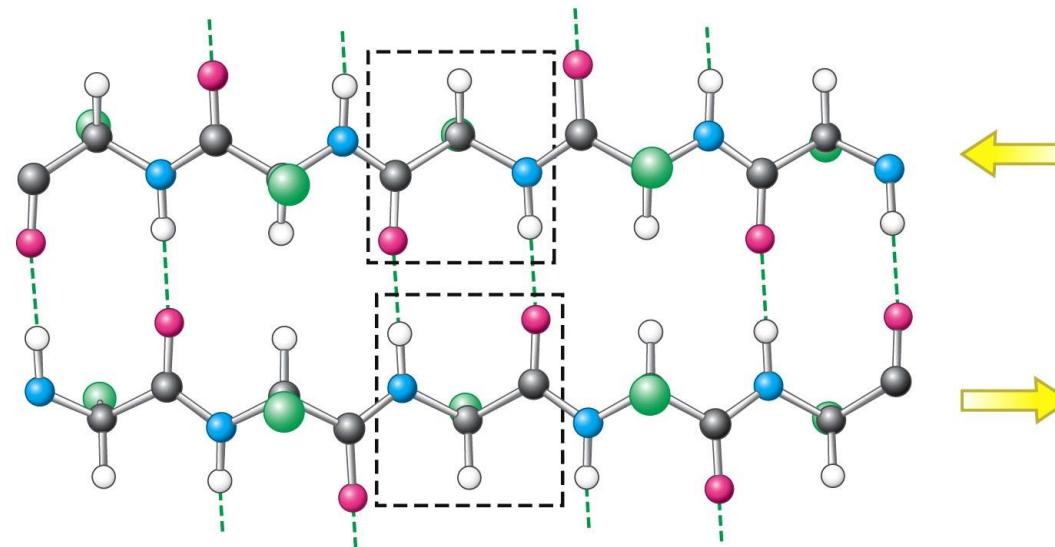


Figure 2.31
Biochemistry, Seventh Edition
© 2012 W. H. Freeman and Company

Protein Structure

Four levels of Protein structure:

1. Primary
2. Secondary
 - 3D arrangement of local segments
 - Hydrogen Bonds
 - Phi/Psi angles
 - “3 State” – H E O

Parallel Beta Sheet

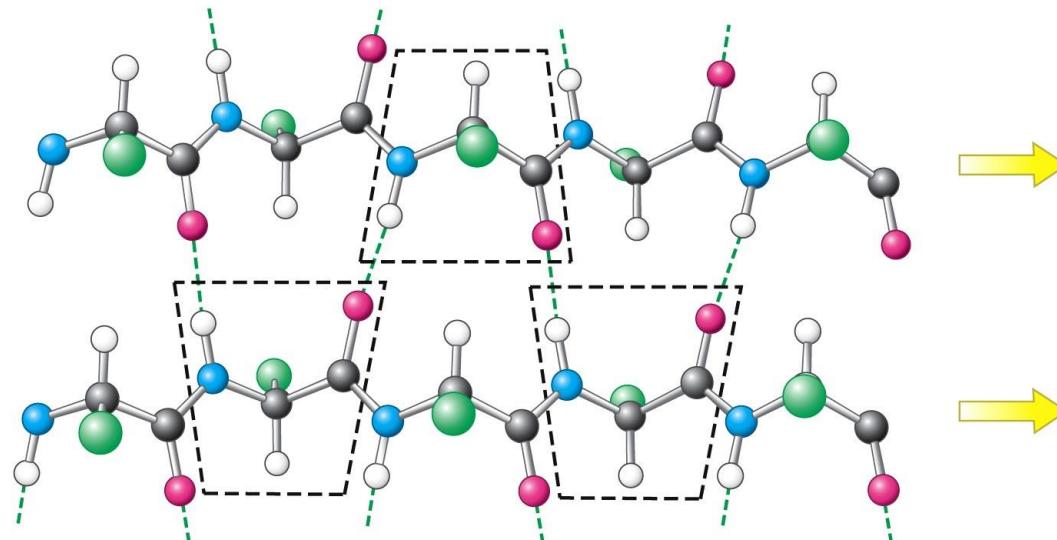


Figure 2.32
Biochemistry, Seventh Edition
© 2012 W. H. Freeman and Company

Protein Structure

Four levels of Protein structure:

1. Primary
2. Secondary

- 3D arrangement of local segments
- Hydrogen Bonds
- Phi/Psi angles
- “3 State” – H E O

Anti-Parallel Beta Sheet

Parallel Beta Sheet

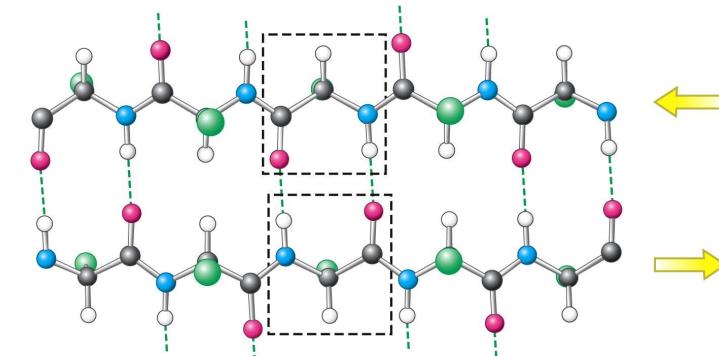


Figure 2.31
Biochemistry, Seventh Edition
© 2012 W. H. Freeman and Company

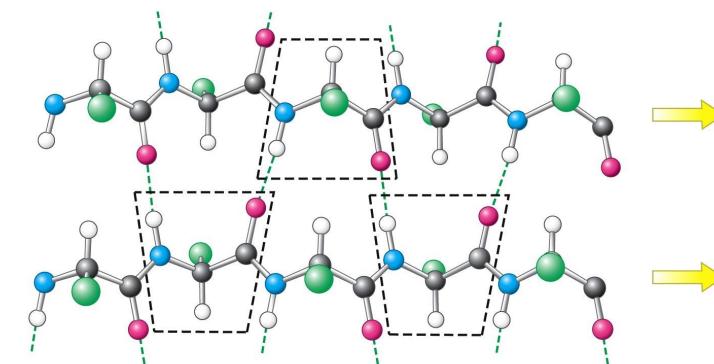


Figure 2.32
Biochemistry, Seventh Edition
© 2012 W. H. Freeman and Company

Protein Structure

Four levels of Protein structure:

1. Primary
2. Secondary
 - 3D arrangement of local segments
 - Hydrogen Bonds
 - Phi/Psi angles

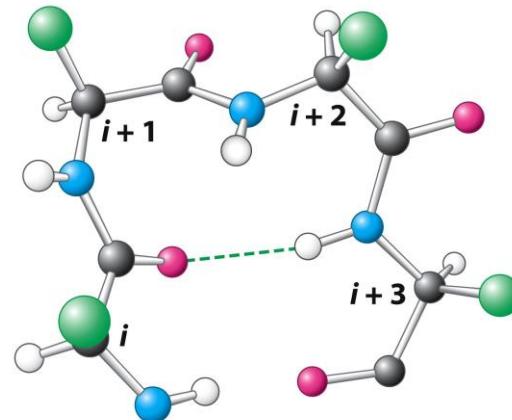
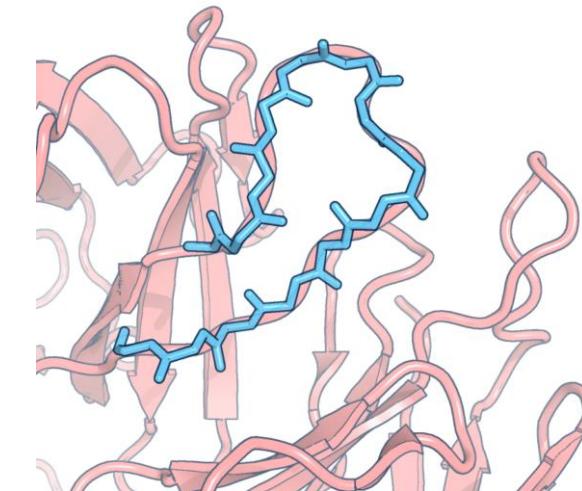


Figure 2.36
Biochemistry, Seventh Edition
© 2012 W. H. Freeman and Company

Beta-turn



Loops

“Other”

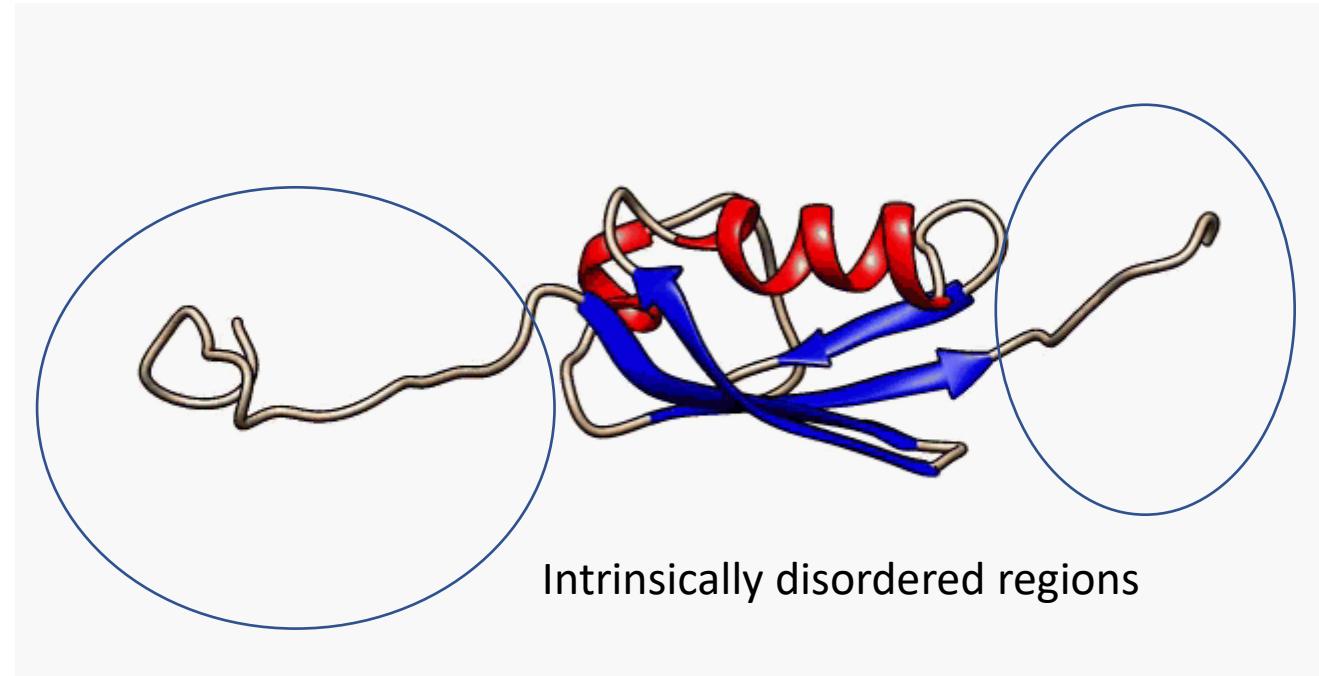
Protein Structure

Four levels of Protein structure:

1. Primary
2. Secondary
 - 3D arrangement of local segments
 - Hydrogen Bonds
 - Phi/Psi angles

Intrinsic Disorder

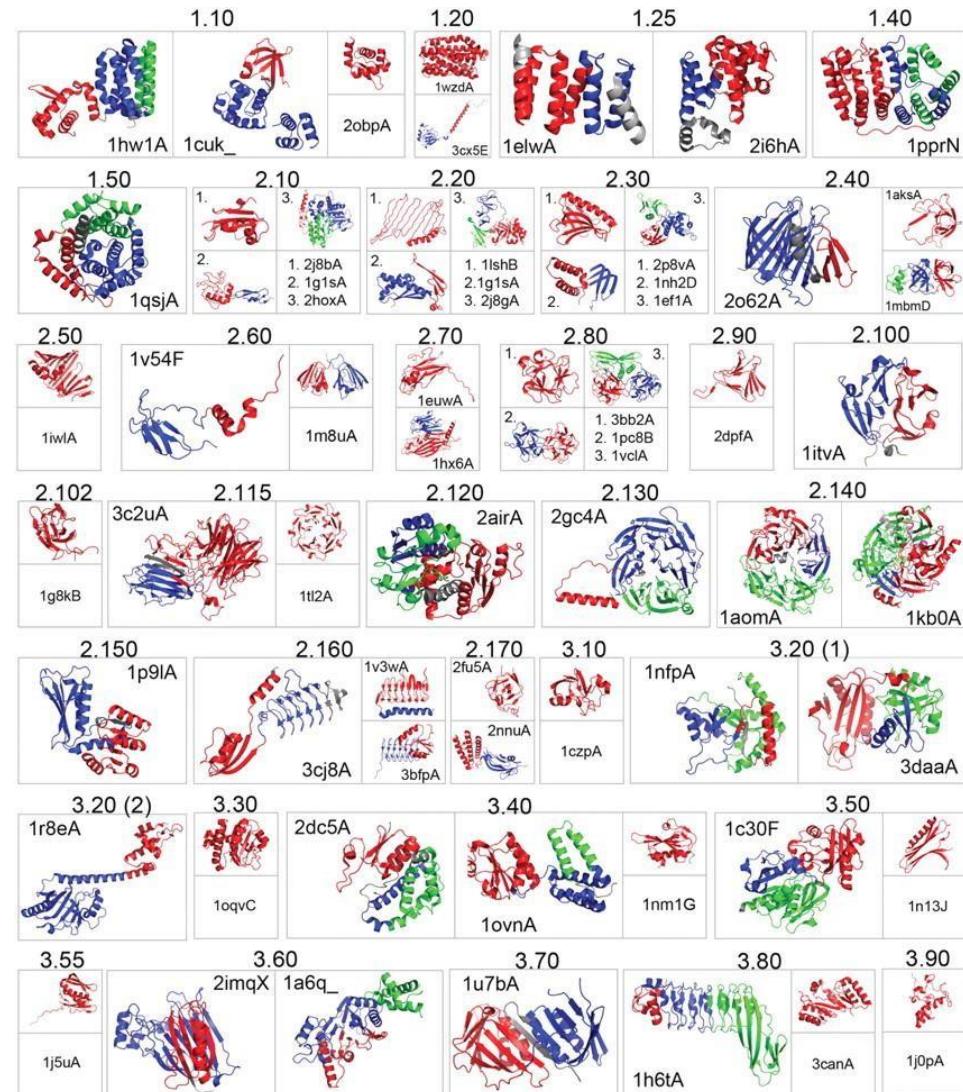
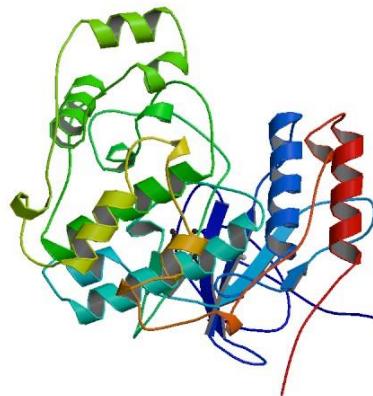
No stable/fixed 3D structure



Protein Structure

Four levels of Protein structure:

1. Primary
2. Secondary
3. Tertiary
 - Protein Fold
 - A single chain

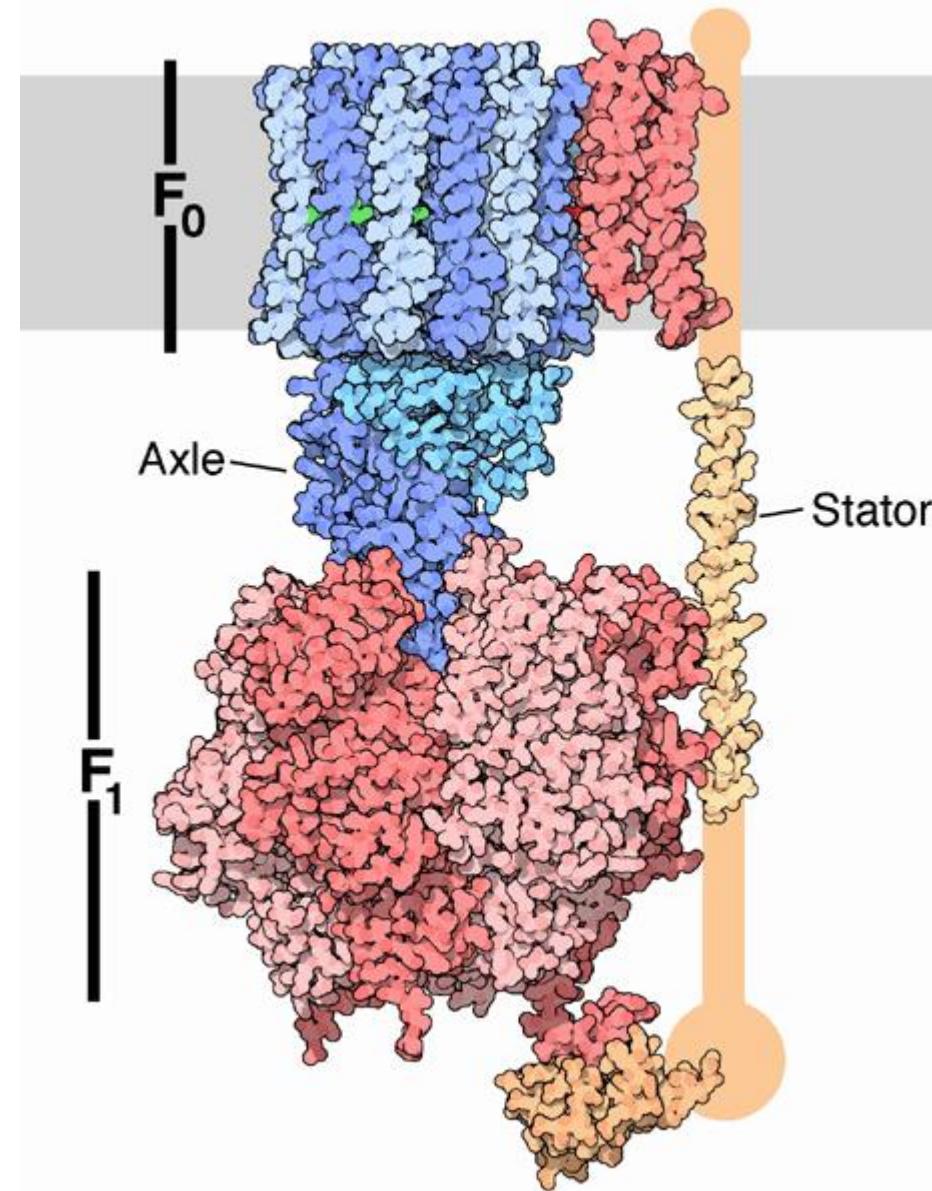


CATH/GENE3D

Protein Structure

Four levels of Protein structure:

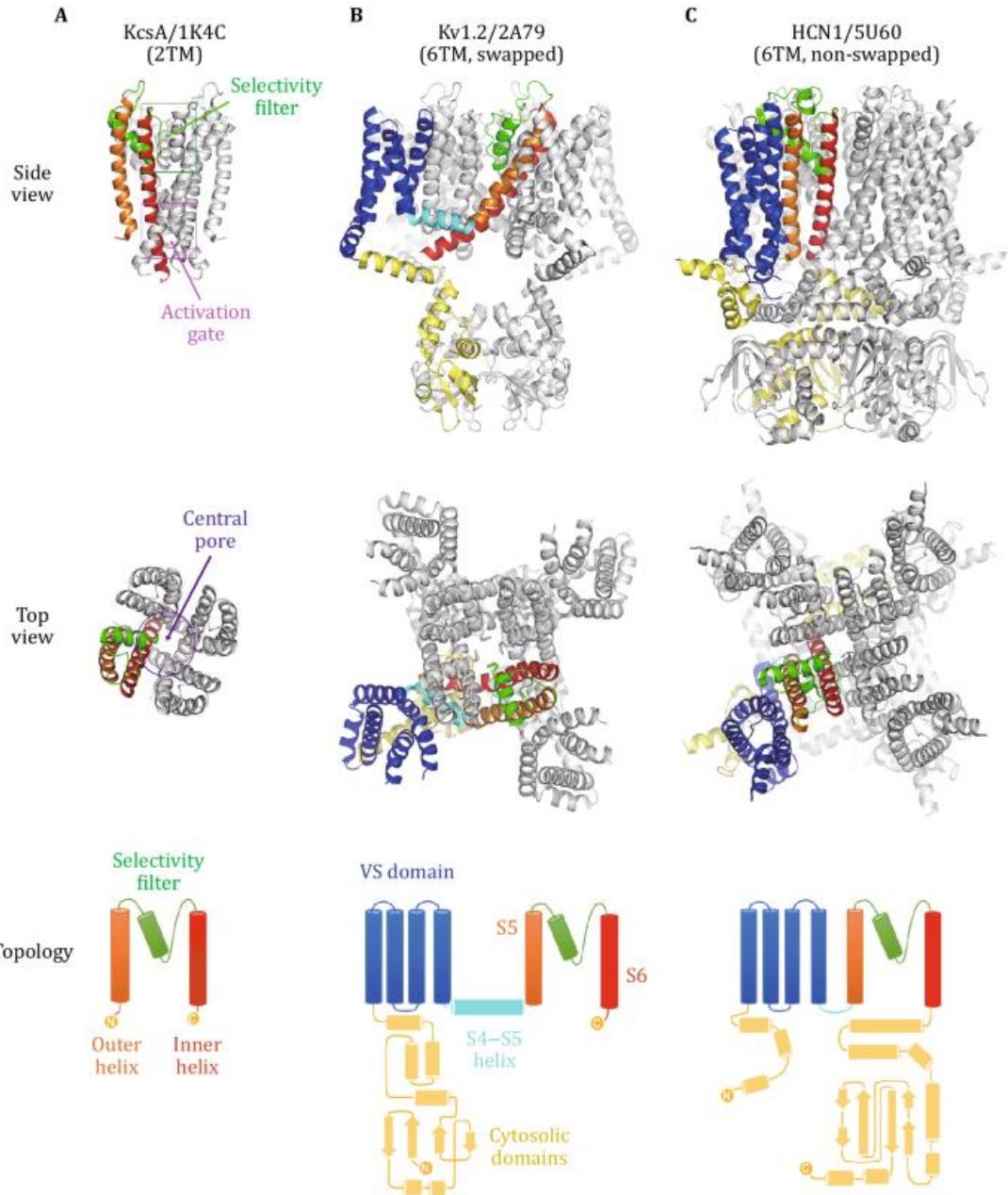
1. Primary
2. Secondary
3. Tertiary
4. Quaternary
 - Arrangement of multiple subunits
 - Protein complexes



Protein Structure

Four levels of Protein structure:

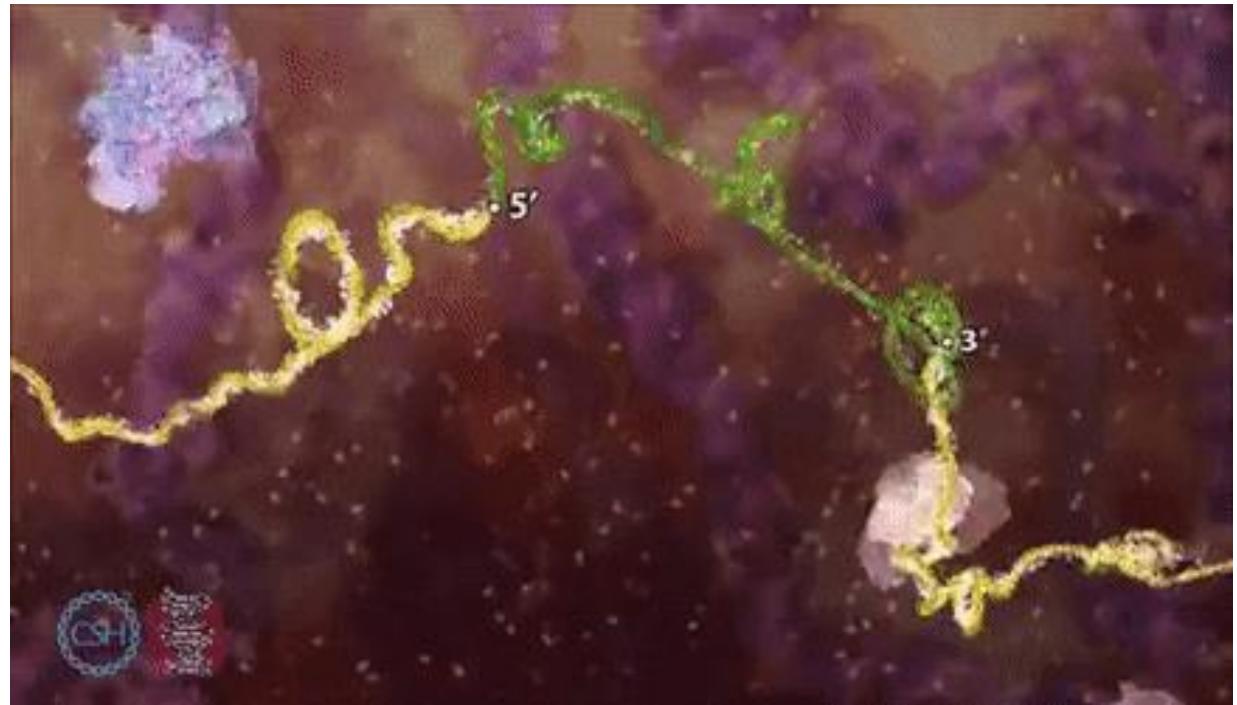
1. Primary
2. Secondary
3. Tertiary
4. Quaternary
 - Arrangement of multiple subunits
 - Protein complexes



Protein Structure

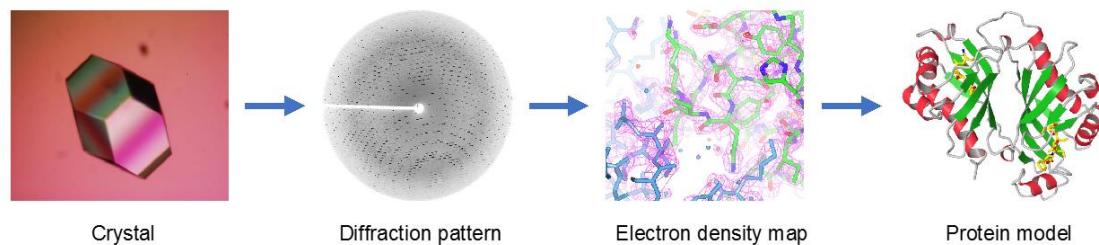
Four levels of Protein structure:

1. Primary
2. Secondary
3. Tertiary
4. Quaternary
 - Arrangement of multiple subunits
 - Protein complexes

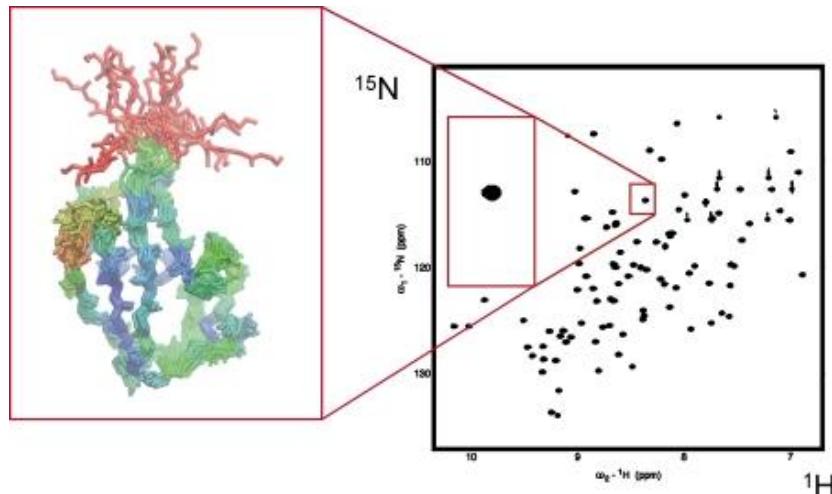


Protein Structure

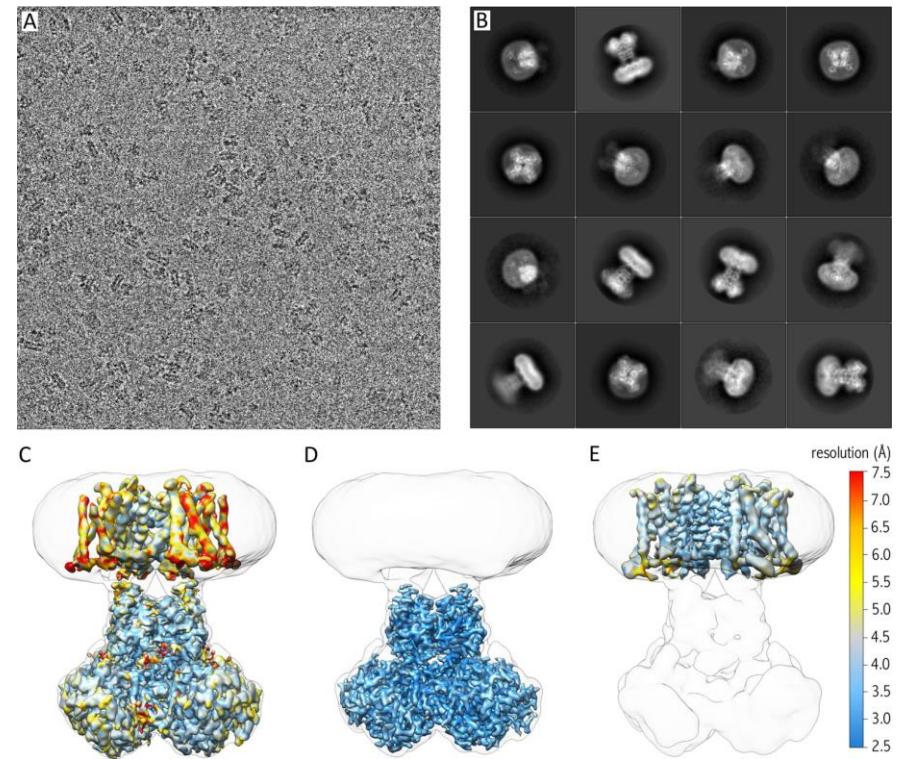
Experimental methods



X-ray crystallography



NMR spectroscopy



Cryo-EM microscopy

<https://pdb101.rcsb.org/>

Structural Data

File formats

- There are two main formats in use:
 - PDB files
 - Older format
 - More human readable
 - Common output of modelling servers

```
ITL  CONSERVATION OF SOLVENT-BINDING SITES IN 10 CRYSTAL FORMS OF T4
ITL 2 LYSOZYME
OMPND MOL_ID: 1;
OMPND 2 MOLECULE: T4 LYSOZYME;
OMPND 3 CHAIN: A;
OMPND 4 EC: 3.2.1.17;
OMPND 5 ENGINEERED: YES
OURCE MOL_ID: 1;
OURCE 2 ORGANISM_SCIENTIFIC: ENTEROBACTERIA PHAGE T4;
OURCE 3 ORGANISM TAXID: 10665;
OURCE 4 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
OURCE 5 EXPRESSION_SYSTEM_PLASMID: M13
EYWD  HYDROLASE(O-GLYCOSYL)
XPDTA X-RAY DIFFRACTION
UTHOR X.-J.ZHANG,B.W.MATTHEWS
EVDAT 3 29-NOV-17 151L 1 HELIX
EVDAT 2 24-FEB-09 151L 1 VERSN
EVDAT 1 30-APR-94 151L 0
RNL AUTH X.J.ZHANG,B.W.MATTHEWS
RNL TITL CONSERVATION OF SOLVENT-BINDING SITES IN 10 CRYSTAL FORMS OF
RNL TITL 2 T4 LYSOZYME.
RNL REF PROTEIN SCI. V. 3 1031 1994
RNL REFN ISSN 0961-8368
RNL PMID 7920248
EMARK 2
```

XYZ coordinates

ATOM	354	U	LEU	A	40	5.561	9.340	-15.975	1.00	51.99	U
ATOM	355	CB	LEU	A	46	6.340	12.024	-14.225	1.00	29.24	C
ATOM	356	CG	LEU	A	46	4.948	12.214	-14.805	1.00	87.60	C
ATOM	357	CD1	LEU	A	46	3.883	11.885	-13.764	1.00	27.41	C
ATOM	358	CD2	LEU	A	46	4.768	13.633	-15.302	1.00	50.09	C
ATOM	359	N	ASP	A	47	7.425	10.348	-16.563	1.00	74.24	N
ATOM	360	CA	ASP	A	47	7.186	9.796	-17.873	1.00	48.30	C
ATOM	361	C	ASP	A	47	7.009	8.310	-17.689	1.00	67.86	C
ATOM	362	O	ASP	A	47	5.949	7.718	-17.911	1.00	69.87	O
ATOM	363	CB	ASP	A	47	8.271	10.149	-18.909	1.00	93.69	C
ATOM	364	CG	ASP	A	47	8.755	11.569	-18.823	1.00	89.47	C
ATOM	365	OD1	ASP	A	47	8.194	12.443	-18.180	1.00	100.00	O
ATOM	366	OD2	ASP	A	47	9.851	11.756	-19.512	1.00	85.74	O
ATOM	367	N	LYS	A	48	8.057	7.722	-17.194	1.00	33.00	N
ATOM	368	CA	LYS	A	48	8.035	6.309	-16.905	1.00	100.00	C
ATOM	369	C	LYS	A	48	6.650	5.888	-16.392	1.00	31.39	C
ATOM	370	O	LYS	A	48	5.972	5.058	-16.994	1.00	67.68	O
ATOM	371	CB	LYS	A	48	9.195	5.906	-15.972	1.00	42.18	C
ATOM	372	CG	LYS	A	48	9.185	4.436	-15.547	1.00	48.87	C
ATOM	373	CD	LYS	A	48	8.719	4.184	-14.105	1.00	100.00	C
ATOM	374	CE	LYS	A	48	7.200	4.033	-13.909	1.00	100.00	C
ATOM	375	NZ	LYS	A	48	6.795	3.218	-12.711	1.00	23.28	N
ATOM	376	N	LYS	A	49	6.200	6.495	-15.206	1.00	39.90	N

Structural Data

File formats

- There are two main formats in use:
 - PDB files
 - Older format
 - More human readable
 - Common output of modelling servers
 - mmCIF files
 - Newer format
 - More computer friendly – nightmare to read
 - BioPython is your friend

```
data_151L
#
_entry.id      151L
#
_audit_conform.dict_name      mmcif_pdbx.dic
_audit_conform.dict_version   5.287
_audit_conform.dict_location  http://mmcif.pdb.org/dictionaries/ascii/mmcif_pdbx.dic
#
loop_
_database_2.database_id
_database_2.database_code
PDB    151L
WWPDB D_1000170119
#
_pdbx_database_status.status_code          REL
_pdbx_database_status.entry_id            151L
_pdbx_database_status.recvd_initial_deposition_date 1994-01-25
_pdbx_database_status.deposit_site        ?
_pdbx_database_status.process_site        BNL
_pdbx_database_status.SG_entry           .
_pdbx_database_status.status_code_sf     REL
_pdbx_database_status.pdb_format_compatible Y
_pdbx_database_status.status_code_mr     ?
_pdbx_database_status.status_code_cs     ?
_pdbx_database_status.methods_development_category ?
#
loop_
_audit_author.name
_audit_author.pdbx_ordinal
'Zhang, X.-J.' 1
'Matthews, B.W.' 2
#
_citation.id
_citation.title
primary
'Cocrystallization of colicin-binding sites in 10 crystal forms of T4 lysiszyme'
```

Structural Data

File formats

- The atomic coordinates of proteins are stored in the Protein Data Bank
- Experimentally determined
 - X-ray crystallography
 - Cryo-EM
 - NMR spectroscopy
- PDB IDs – “XXXX”, 4 letters and numbers

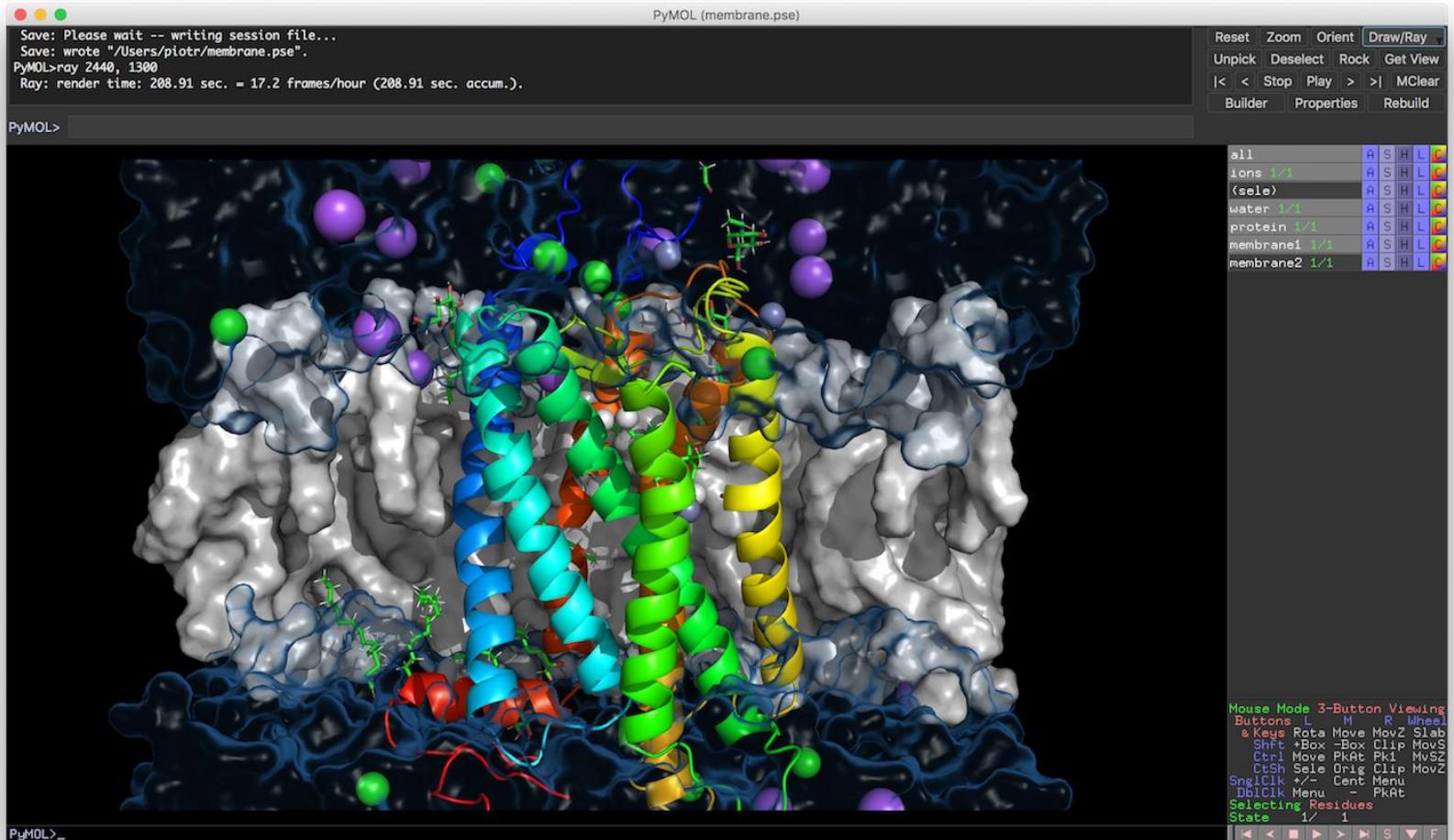


Protein Structure

<https://pymol.org/edu/?q=educational/>

File formats

- Can open both in protein visualisation software
- Or read them in a plain text editor



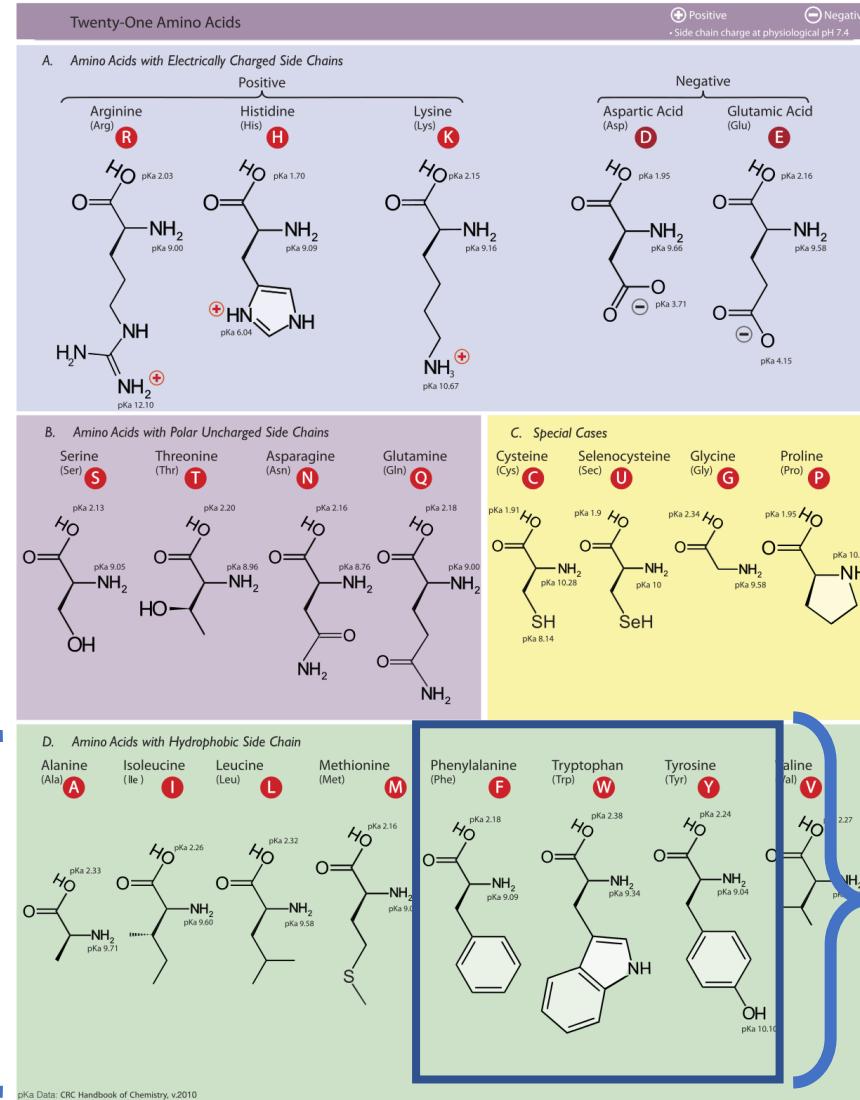
Seq to Structure

How to do we go from sequence to structure?

Sequence gives us:

1. Type of residue

Hydrophobic

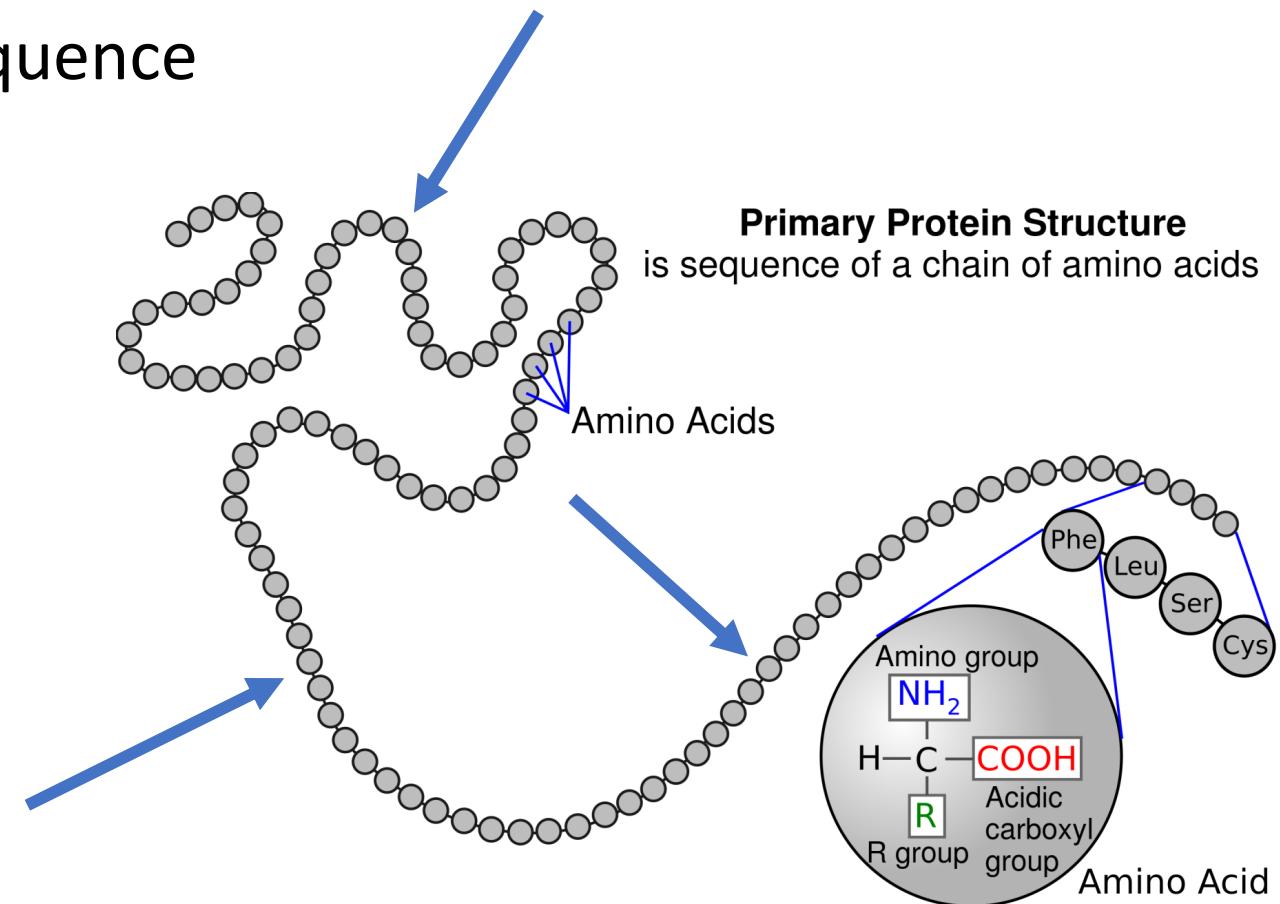


Seq to Structure

How to do we go from sequence to structure?

Sequence gives us:

1. **Type of residue**
2. **Position of residue**



Seq to Structure

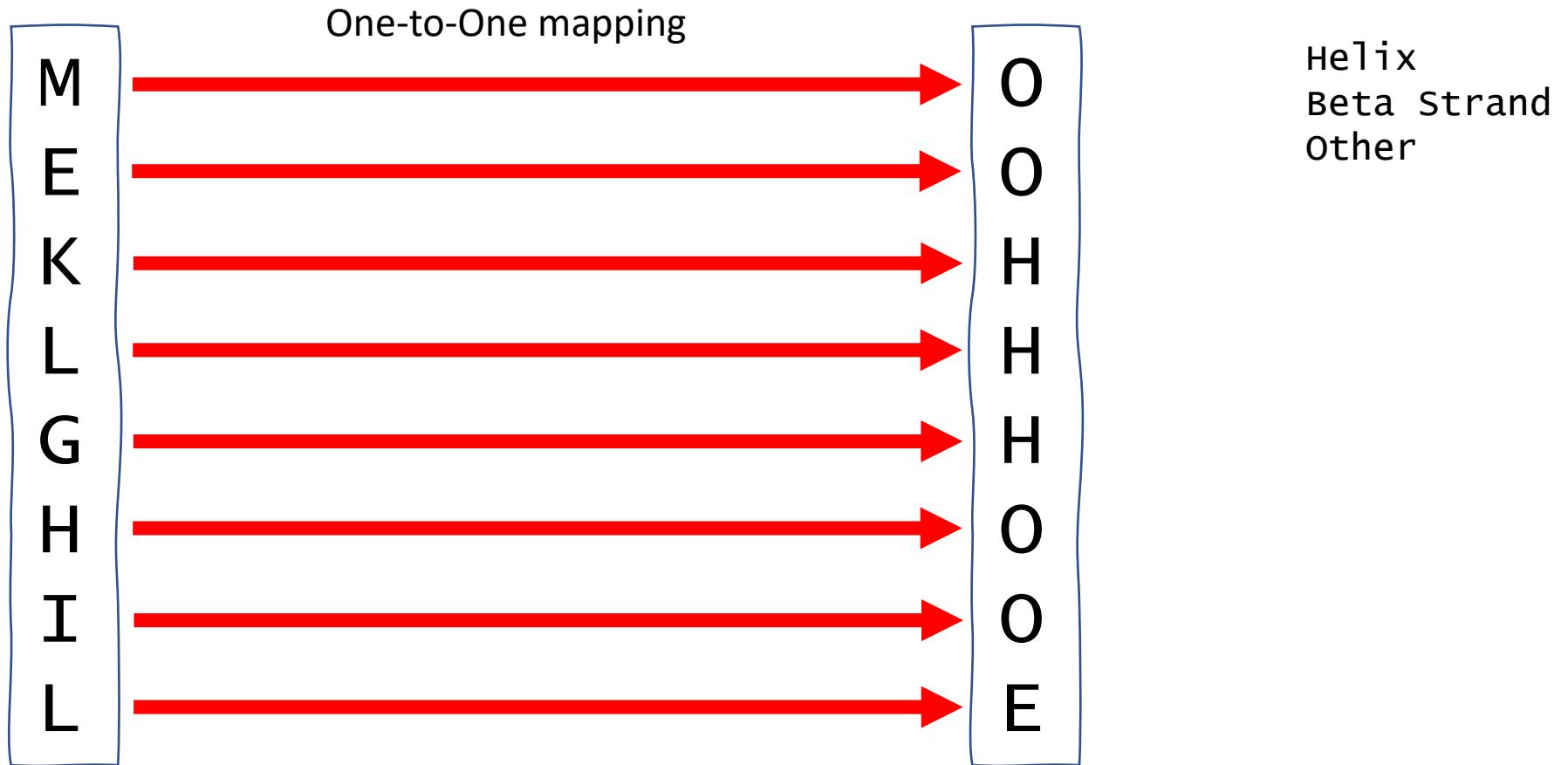
Predicting Secondary structure:

M
E
K
L
G
H
I
L

H Helix
E Beta Strand
O other

Seq to Structure

Predicting Secondary structure:



Seq to Structure

Predicting Secondary structure:

- Simple version – AA propensities from protein structures
 - Poor accuracy
 - No position information
- Example - **Chou and Fasman**

Amino acid	Frequency	P_α	P_β	P_c
A—Ala	7.73	1.39	0.75	0.8
C—Cys	1.84	0.74	1.31	1.05
D—Asp	5.82	0.89	0.55	1.33
E—Glu	6.61	1.35	0.72	0.86
F—Phe	4.05	1.01	1.43	0.76
G—Gly	7.11	0.47	0.65	1.62
H—His	2.35	0.92	0.99	1.07
I—Ile	5.66	1.04	1.71	0.59
K—Lys	6.27	1.11	0.83	1
L—Leu	8.83	1.32	1.1	0.68
M—Met	2.08	1.21	0.99	0.83
N—Asn	4.5	0.77	0.62	1.39
P—Pro	4.52	0.5	0.44	1.72
Q—Gln	3.94	1.29	0.76	0.89
R—Arg	5.03	1.17	0.91	0.91
S—Ser	6.13	0.82	0.85	1.24
T—Thr	5.53	0.76	1.23	1.07
V—Val	6.91	0.89	1.86	0.64
W—Trp	1.51	1.06	1.3	0.79
Y—Tyr	3.54	0.95	1.5	0.78

Seq to Structure

Predicting Secondary structure:

- Integration of position information – “sliding windows”
- Scoring matrices
- Example – **GOR algorithm**



A B C D E F G H I J K L M N O P Q R S T U V W X Y Z



A B C D E F G H I J K L M N O P Q R S T U V W X Y Z



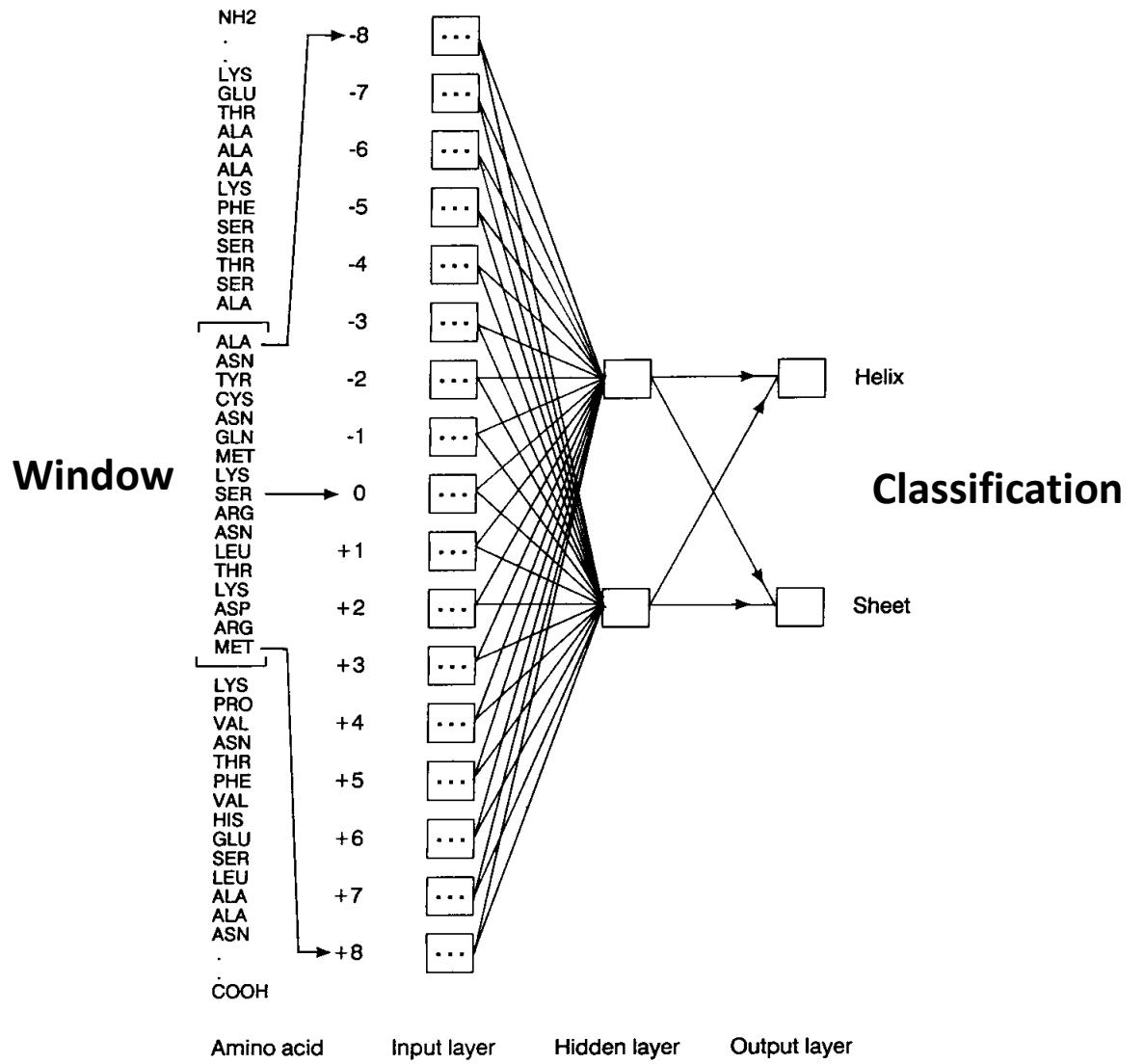
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Seq to Structure

Predicting Secondary structure:

- Machine learning algorithms
 - Neural Networks
 - Support vector machines
 - Random Forest

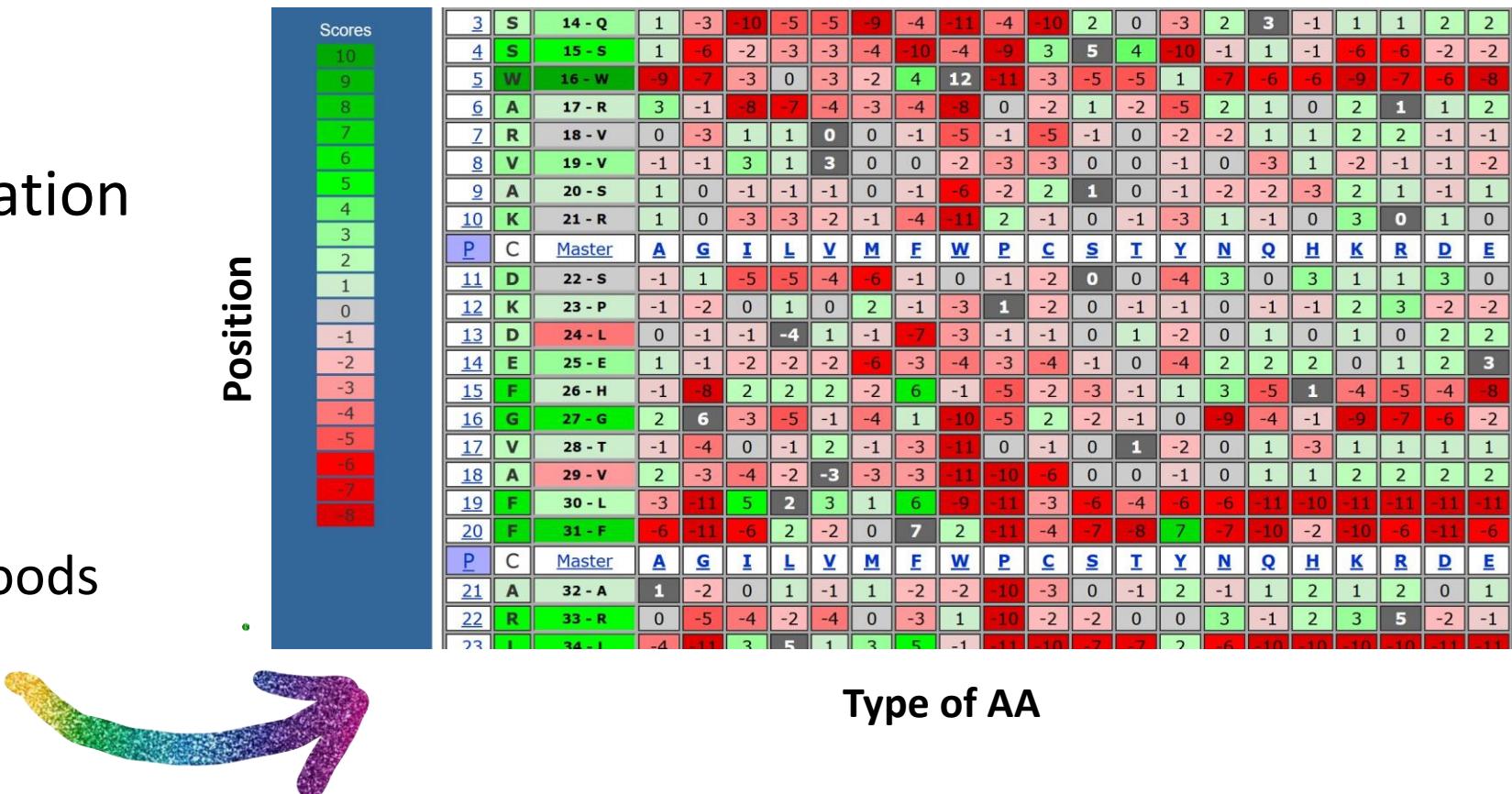
Biophysics: Holley and Karplus



Seq to Structure

Predicting Secondary structure:

- Evolutionary Information
 - Multiple sequence alignments
 - Profiles – probabilities/likelihoods
 - PSSMs – **PSIBLAST**



Seq to Structure

Predicting Secondary structure:

- Nowadays, **Deep learning**
 - **Evolutionary information**
 - Prediction of contacts
 - Prediction of Phi/Psi angles
 - Prediction of SASA
- **~85% accuracy**

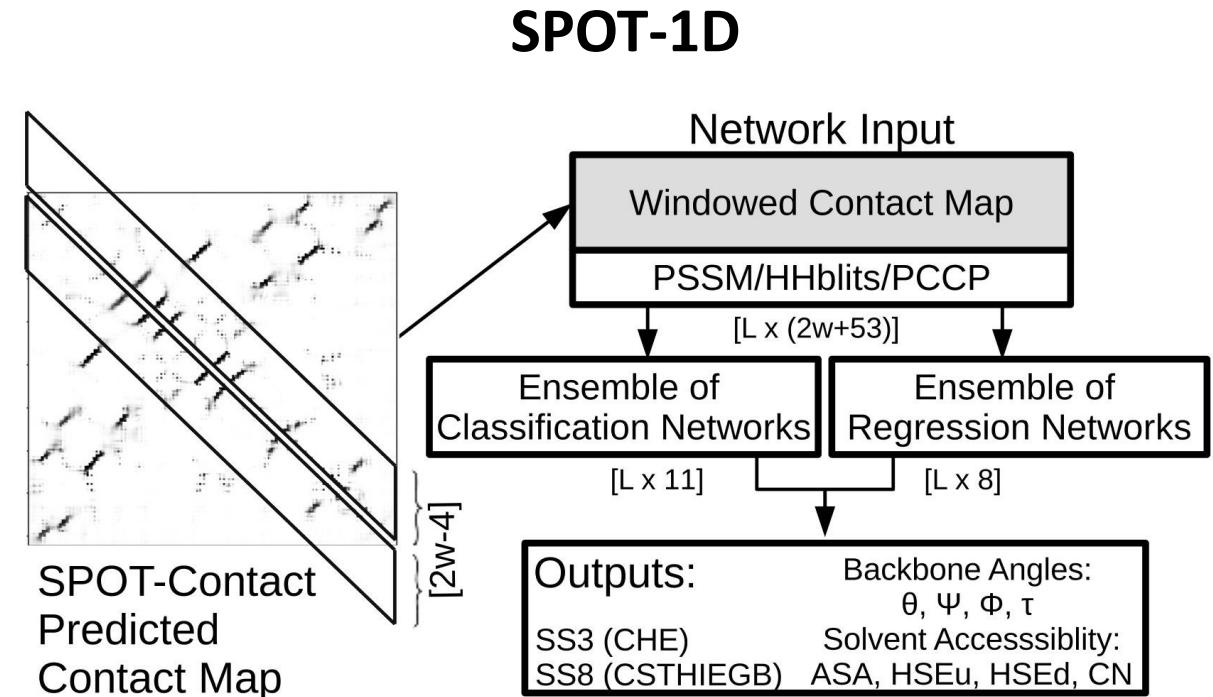
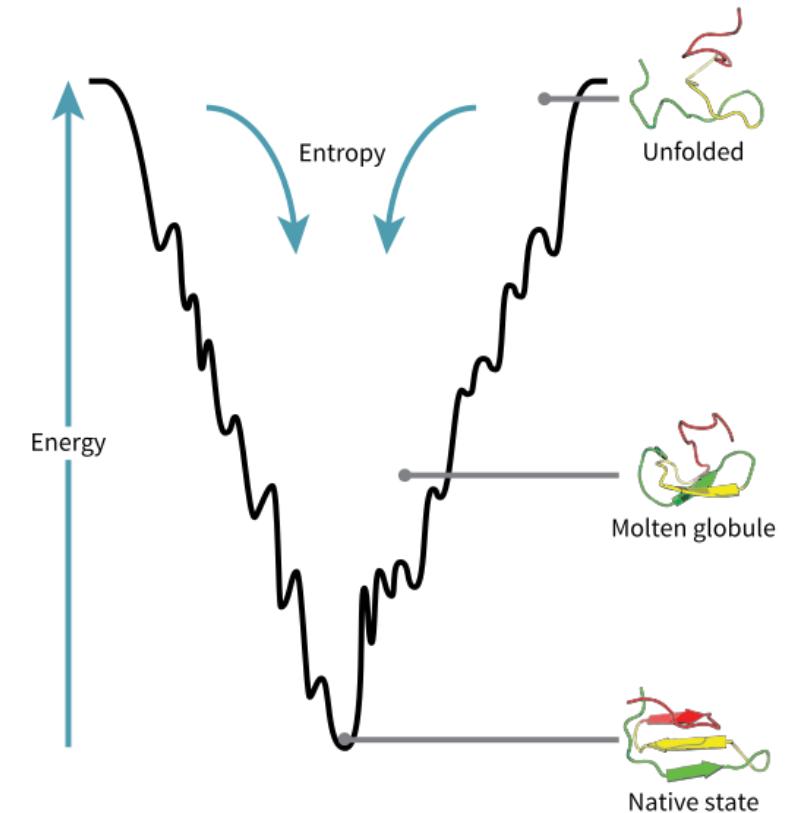


Figure S1: The overview of the SPOT-1D architecture for a target protein of length L and a window size of w .

Hanson et al. 2019

Seq to Structure

- How do we predict **tertiary** structure or fold?
- Difficulties:
 - Search space is **MASSIVE**
 - Levinthal's Paradox (1960s) – time to sample >>> age of universe – **POWER LAW**
 - Incomplete understanding of physics of protein folding
 - Protein-solvent interactions exhibit quantum-wavelike behaviour
 - Computational costs
- Need to reduce search space



Seq to Structure

Structure is **MORE CONSERVED** than sequence

DYDLKFGMNAGTSSNEYKAAEMFAKEVKEKSQGKIEISLYPSSQLGDDRAMLKQLKDGSLDFTFAESARF 2cex_1
VTWRLASSFPKSLDTIFGGAEVLSKXLSEATDGNFQIQVFSAGELVPGLQAADAVTEGTVECCHTVGYYY 2hzk_2

QLFYPEAAVFALPYVISNYNVAQKALFDTEFGKDLIKKMDKDLGVTL LSQAYNGTRQTTSNRAINSIADM 2cex_1
WGKDPTFALAAAVPFSLSARGINAWHYHGGIDLYNEFLSQHNIVAFPGGNTGVQXGGWFRREINTVADX 2hzk_2

KGLKLRVPNAATNLAYAKYVGASPTPMFSEVYLALQTNAVDGQENPLAAVQAQKFYEVQKFLAMTNHIL 2cex_1
QGLKXRVGGFAGKVXERLGVVPPQQIAGGDIYPALEKGTIDATEWVGPYDDEKLGFFKVAPYYYYPGWWEG 2hzk_2

NDQLYLVSNETYKELPEDLQVKVVKDAEENA AKYHTKL FVDGEKDLVTFFEKQGVKITHPDLVPFKESMKP 2cex_1
GPTVHFxFNKSAYEGLPTYQSLLRTACHAADANXLQLYDWKNPTAIKSLVAQGTQLRPFSPPEILQACFE 2hzk_2

YYAEFKQTGQKGESALKQIE 2cex_1
AANEVYAXEXEASNPALKKIWD 2hzk_2

Seq to Structure

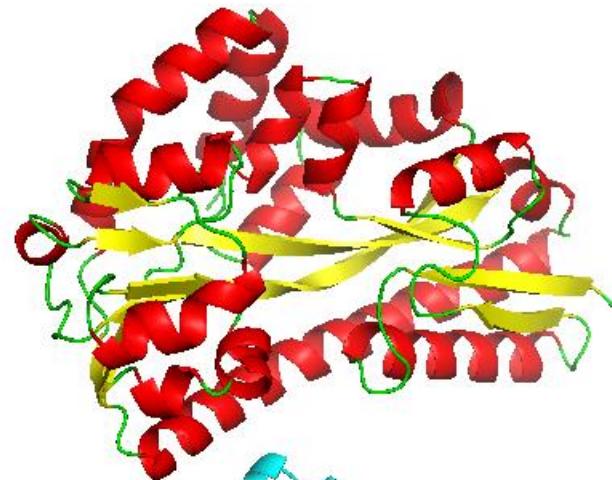
Structure is **MORE CONSERVED** than sequence

DYDLKFGMNAGTSSNEYKAAEMFAKEVKEKSQGKIEISLYPSSQLGDDRAMLKQLKDGSLDFTAESARF	2cex_1	
VTWRLASSFPKSLDTIFGGAEVLSKXLSEATDGNFQIQVFSAGELVPGLQAADAVTEGTVECCHTVYYY	2hzk_2	
*** * * * *	*	
QLFYPEAAVFALPYVISNYNVAQKALFDTEFGKDLIKKMDKDLGVTL LSQAYNGTRQTTSNRAINSIADM	2cex_1	
WGKDPTFALAAAVPFSLSARGINAWHYHGGIDLYNEFLSQHNIVAFPGGNTGVQXGGWFRREINTVADX	2hzk_2	
* * *	*	** **
KGLKLRVPNAATNLAYAKYVGASPTPMFSEVYLALQTNAVDGQENPLAAVQAQKFYEVQKFLAMTNHIL	2cex_1	
QGLKXRVGGFAGKVXERLGVVPPQQIAGGDIYPALEKGTIDATEWVGPYDDEKLGFFKAPYYYYPGWWEG	2hzk_2	
*** ** * * *	* * *	*
NDQLYLVSNETYKELPEDLQVKVKDAEENA AKYHTKL FVDGEKDLVTFFEKGVKITHPDLVPKESMKP	2cex_1	
GPTVHFxFNKSAYEGLTPYQSLLRTACHAADANXLQLYDWKNPTAIKSLVAQGTQLRPFSSPEILQACFE	2hzk_2	
* * * * *	* * *	*
YYAEFKQTGQKGESALKQIE	2cex_1	
AANEVYAXEXEASNPALKKIWD	2hzk_2	
*	*	

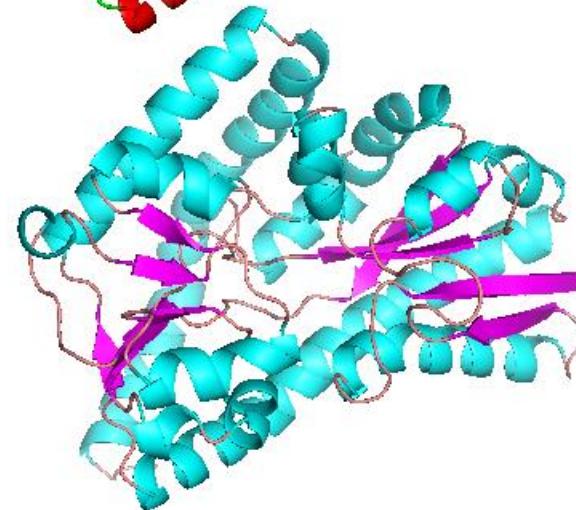
Exact matches = 35/ 301 = 11%

Seq to Structure

Structure is **MORE CONSERVED** than sequence



2cex



2hzk

Exact matches = $35 / 301 = 11\%$

Seq to Structure

Template-based structure prediction

- Homology Modelling
- Protein Threading

De novo/Template-free structure prediction

- Physics based
- Knowledge-based

Template-based

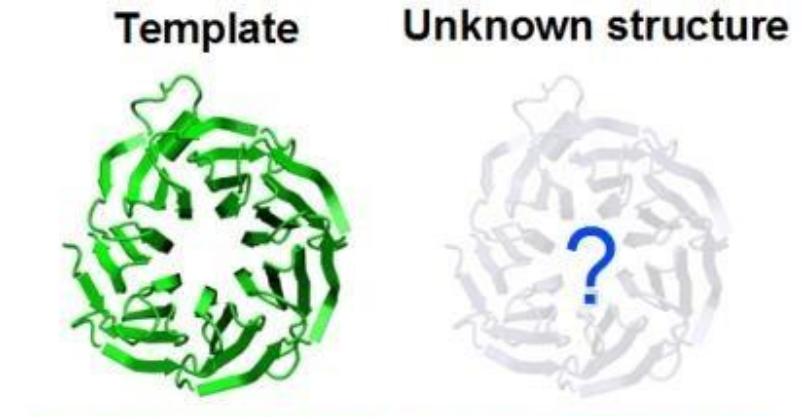
Unknown structure



- Homology modelling
 - Proteins that have similar sequence will probably have similar structures

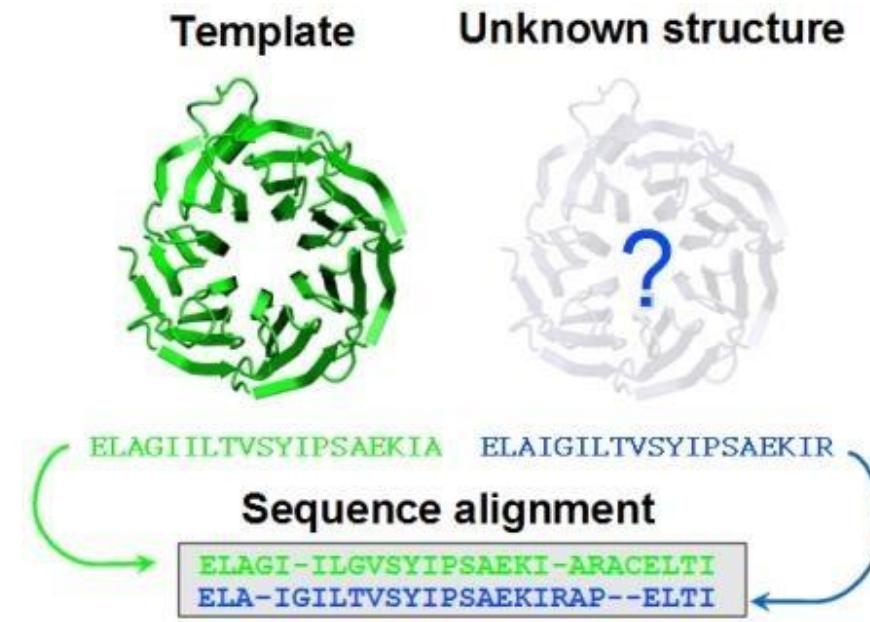
Template-based

- Homology modelling
 - Proteins that have similar sequence will probably have similar structures
 - Find most similar protein with a solved structure – **TEMPLATE**



Template-based

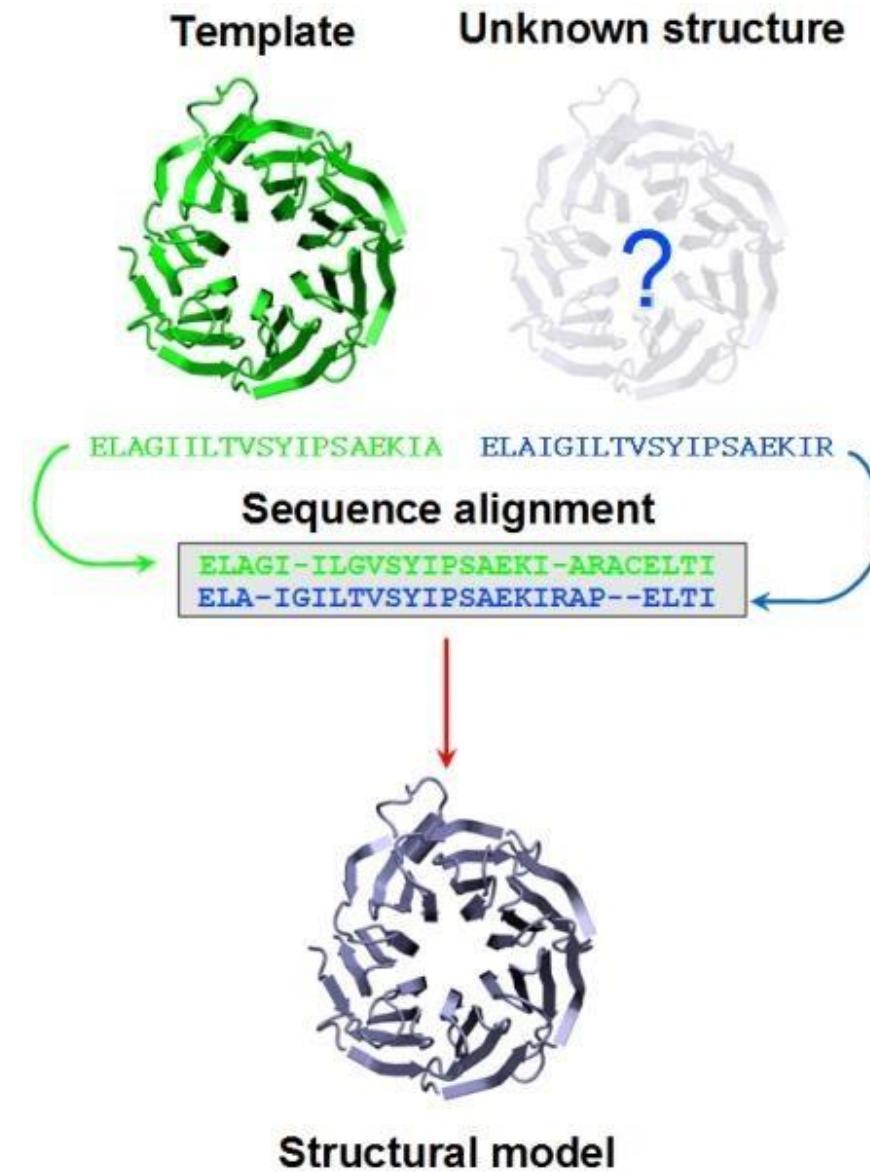
- Homology modelling
 - Proteins that have similar sequence will probably have similar structures
 - Find most similar protein with a solved structure – **TEMPLATE**
 - **Sequence-Sequence** alignment



Template-based

- Homology modelling

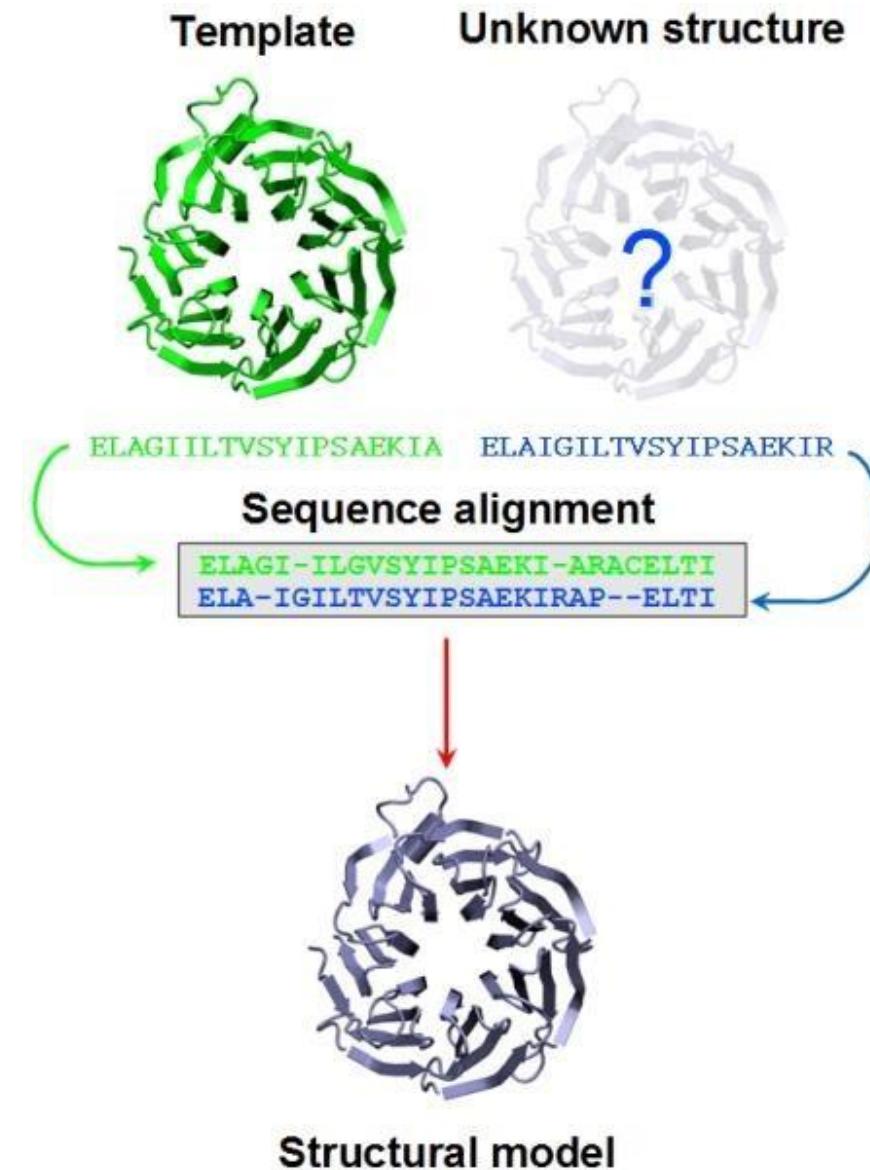
- Proteins that have similar sequence will probably have similar structures
- Find most similar protein with a solved structure – **TEMPLATE**
- **Sequence-Sequence** alignment
- Replace residues in template with those from target



Template-based

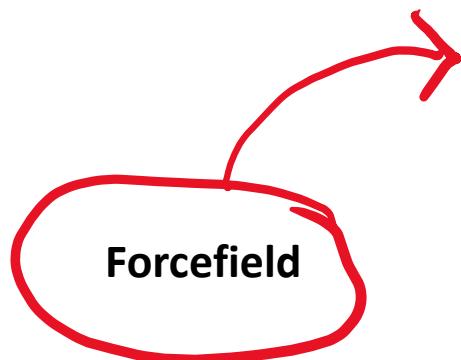
- Homology modelling

- Proteins that have similar sequence will probably have similar structures
- Find most similar protein with a solved structure – **TEMPLATE**
- **Sequence-Sequence** alignment
- Replace residues in template with those from target
- >30-40% sequence identity rule-of-thumb as minimum requirement
- Example – **SWISS-MODEL**

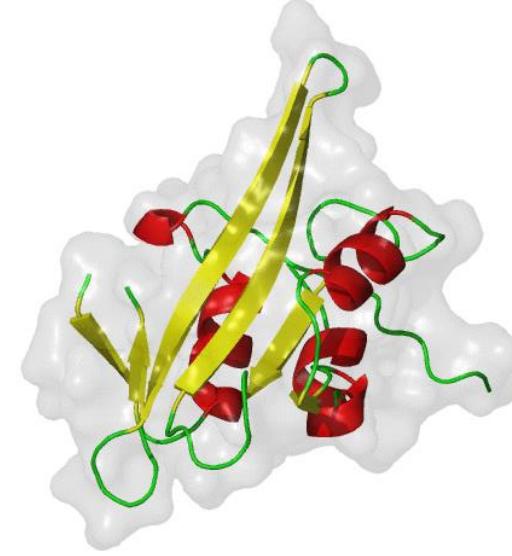


Model Refinement

- Models can be made more accurate
- Final “polish” to get even closer to native state
- Physics based methods e.g. **molecular dynamics**



$$\begin{aligned} U(R) = & \sum_{bonds} k_r (r - r_{eq})^2 \\ & + \sum_{angles} k_\theta (\theta - \theta_{eq})^2 \\ & + \sum_{dihedrals} k_\phi (1 + \cos[n\phi - \gamma]) \\ & + \sum_{impropers} k_\omega (\omega - \omega_{eq})^2 \\ & + \sum_{atoms} \epsilon_{ij} \left[\left(\frac{r_m}{r_{ij}} \right)^{12} - 2 \left(\frac{r_m}{r_{ij}} \right)^6 \right] \\ & + \sum_{i < j}^{atoms} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \end{aligned}$$



bond

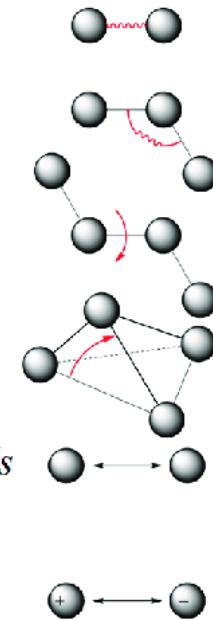
angle

dihedral

improper

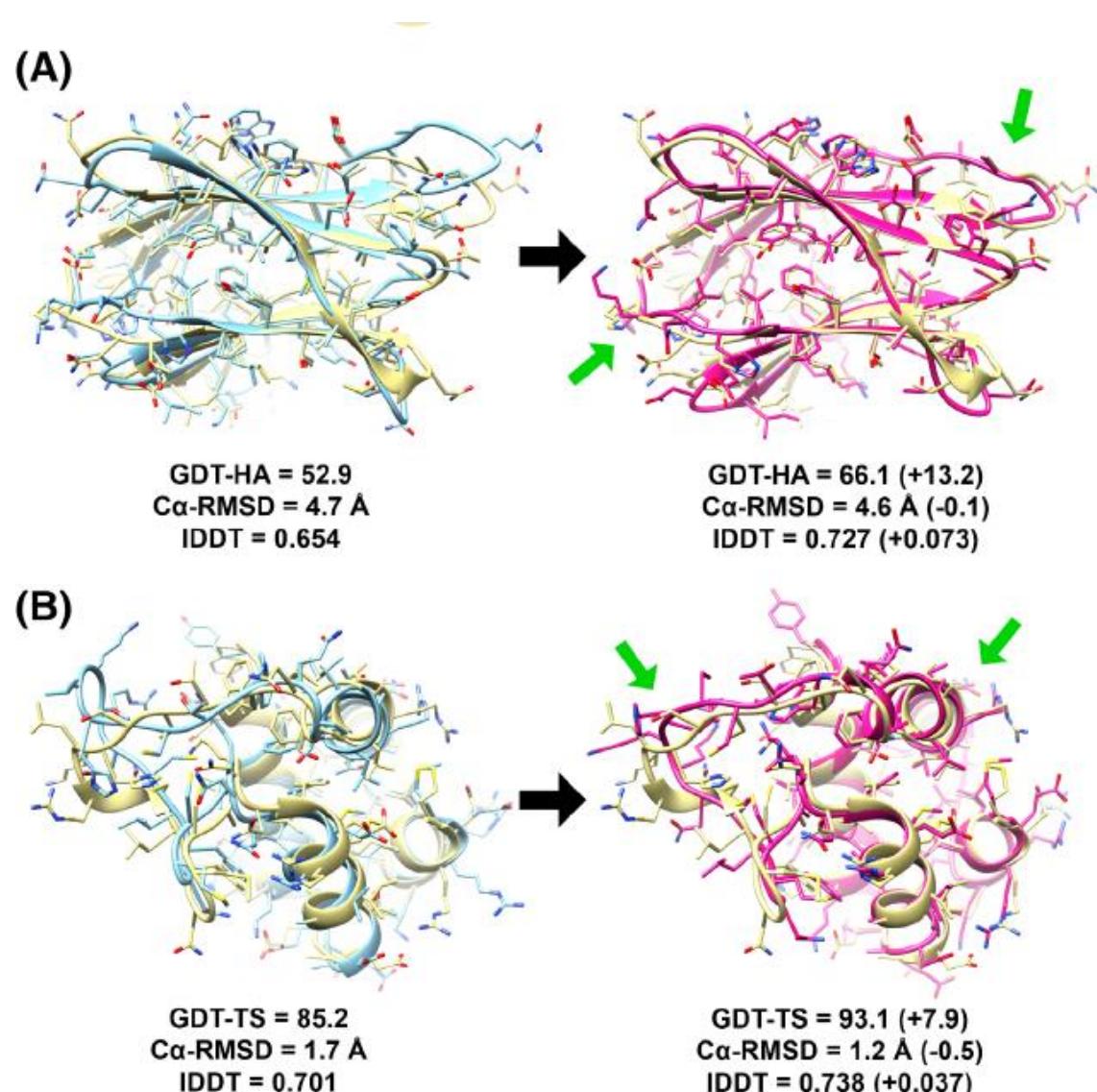
van der Waals

electrostatic



Model Refinement

- Models can be made more accurate
- Final “polish” to get even closer to native state
- Physics based methods e.g. **molecular dynamics**
- Example - PREFMD



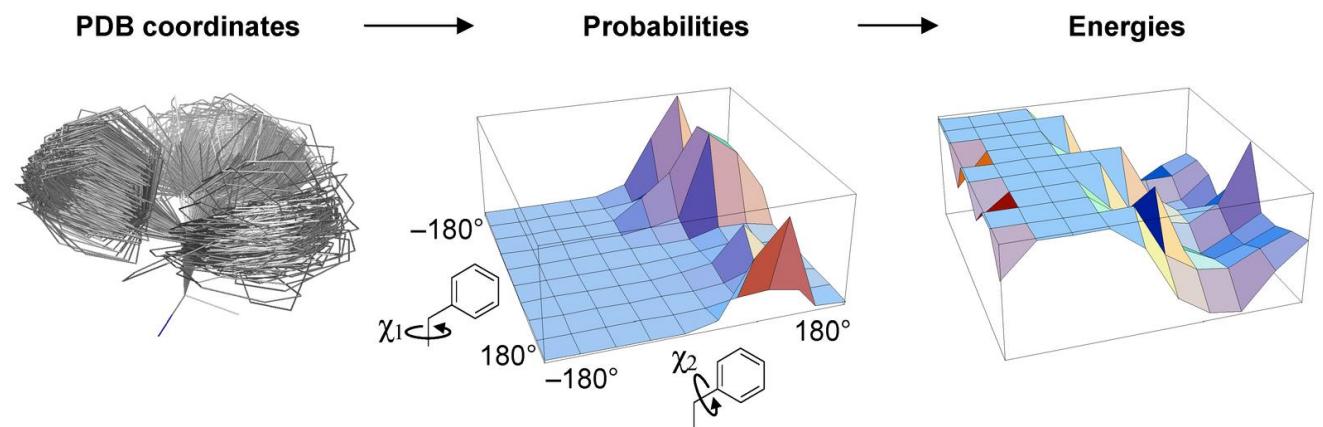
Model Refinement

- Models can be made more accurate
- Final “polish” to get even closer to native state

- **Knowledge-Based/Statistical Potentials**

- **Example - Rosetta**

“...scoring functions derived from an analysis of known protein structures in the Protein Data Bank (PDB).”



Chi-angle probability distribution for a sidechain – More probable states are probably lower energy states.

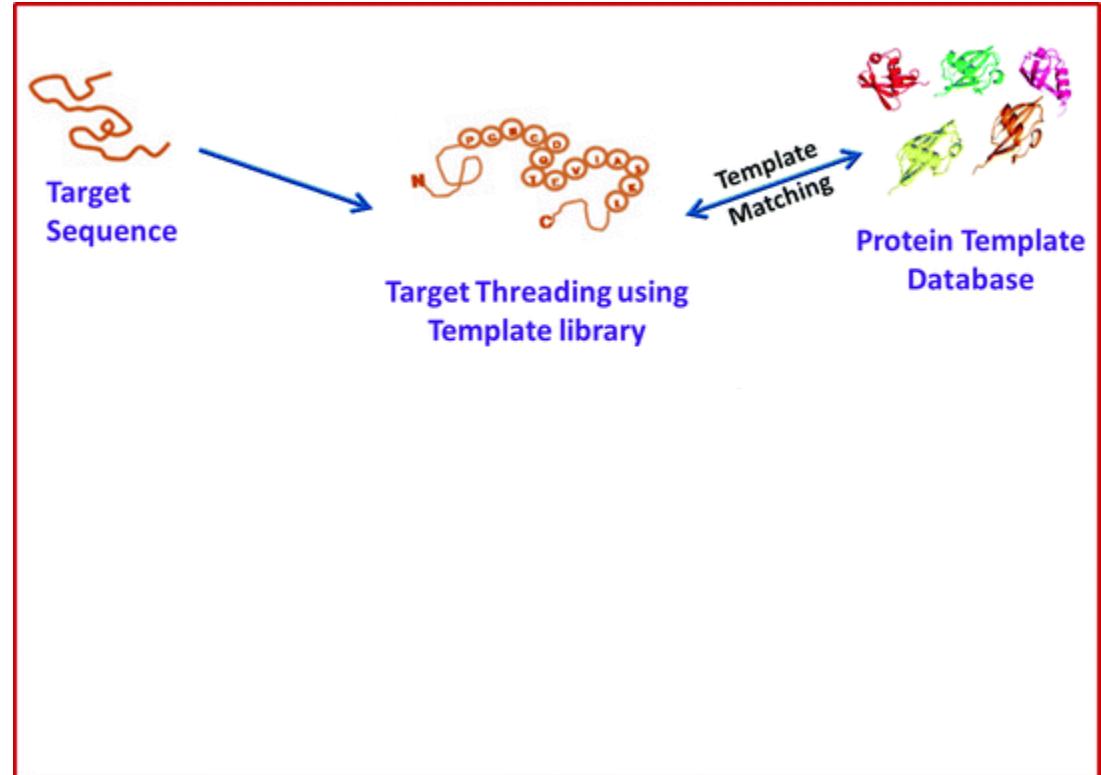
Template-based

- Protein Threading
 - **Sequence-fold alignment**
 - If close homologues don't exist



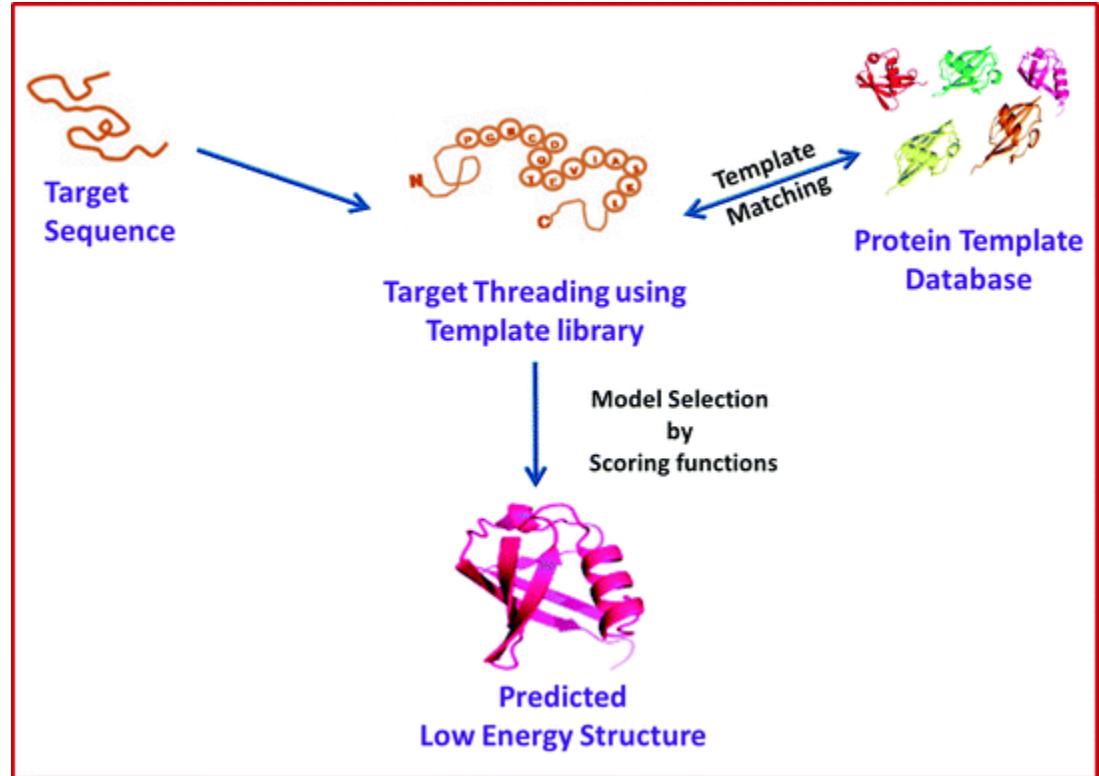
Template-based

- Protein Threading
 - **Sequence-fold alignment**
 - If close homologues don't exist
 - “**thread**” input sequence onto known protein folds



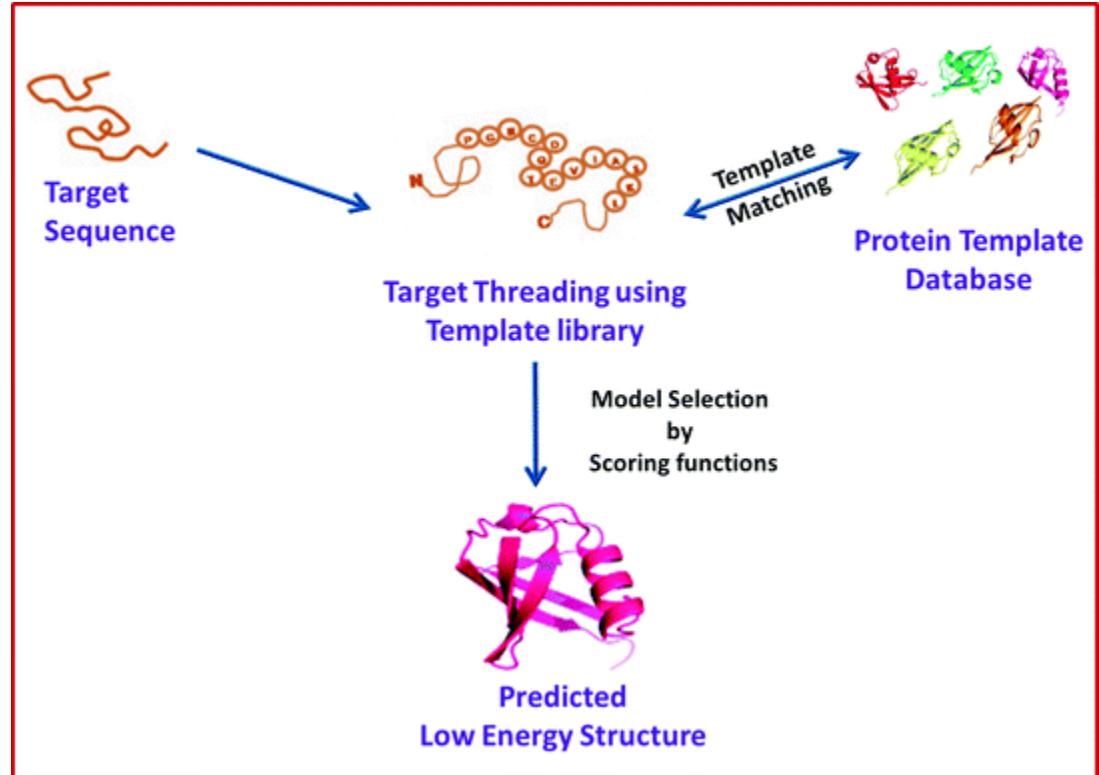
Template-based

- Protein Threading
 - **Sequence-fold alignment**
 - If close homologues don't exist
 - “**thread**” input sequence onto known protein folds
 - Evaluate resulting models using **statistical or physics based scoring functions**



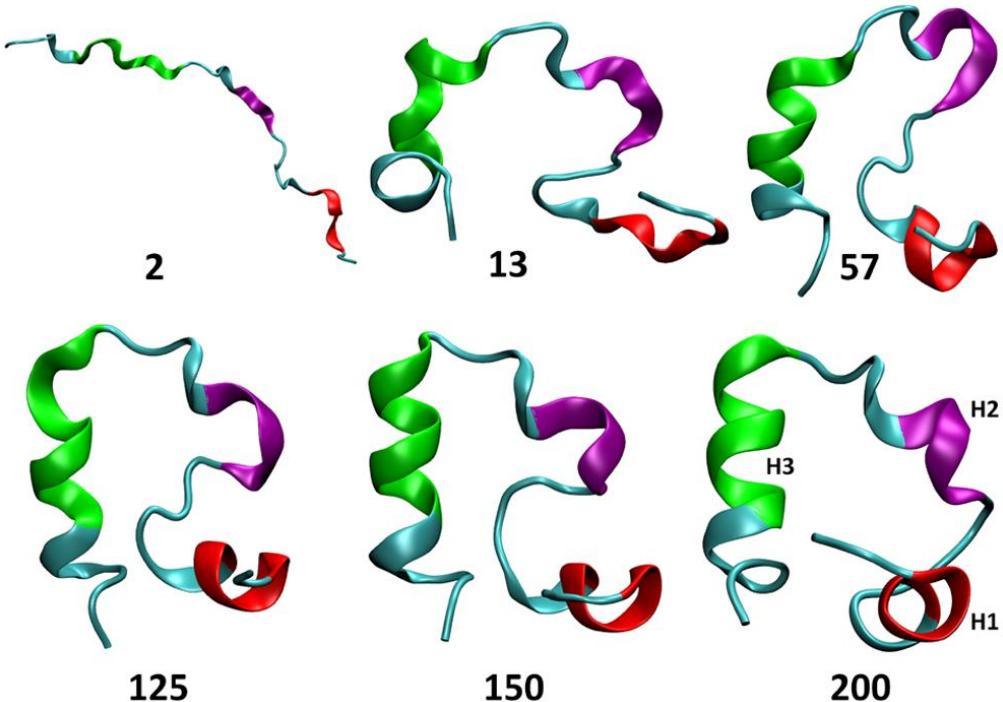
Template-based

- Protein Threading
 - Works because **fold space is limited**
 - 90% of new structures have a fold that already exists in the PDB
- Example - ITASSER



Template-free

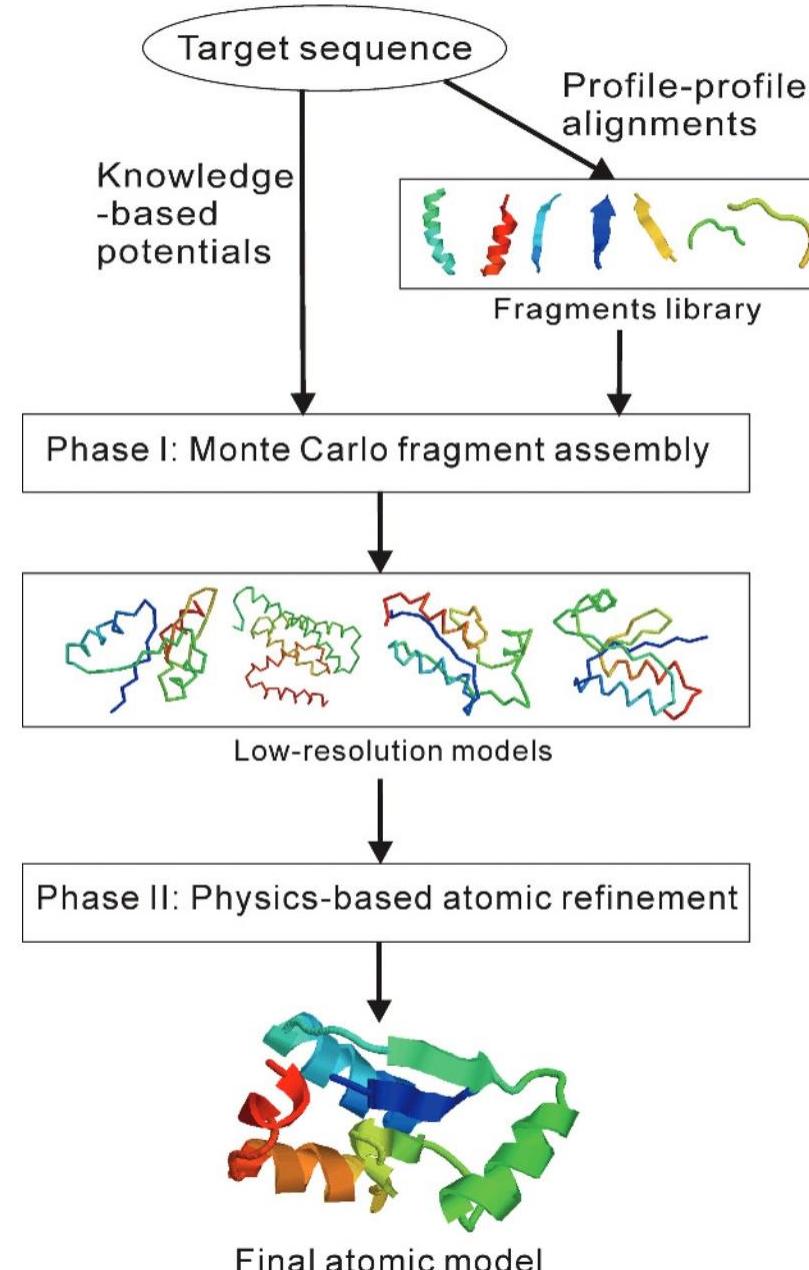
- *Ab Initio* or “*de novo*”
 - When you have no homologous/fold template
 - Physics based methods
 - E.g. Molecular Dynamics – Anton, D.E. Shaw; Folding@home



Template-free

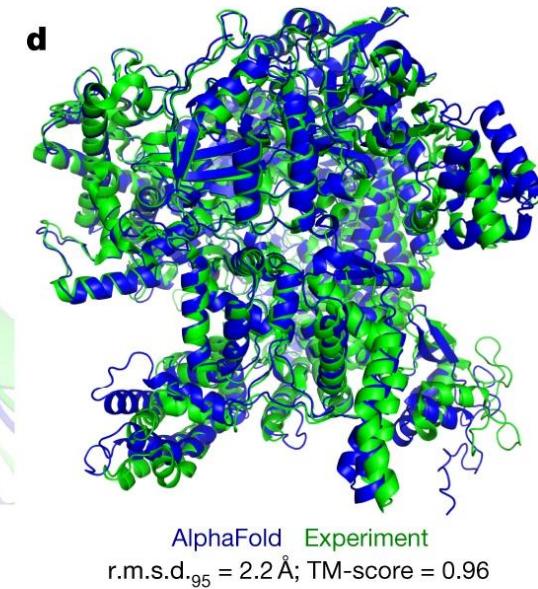
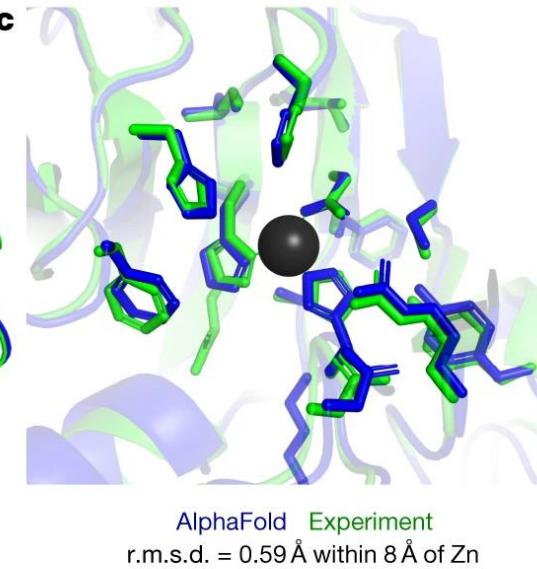
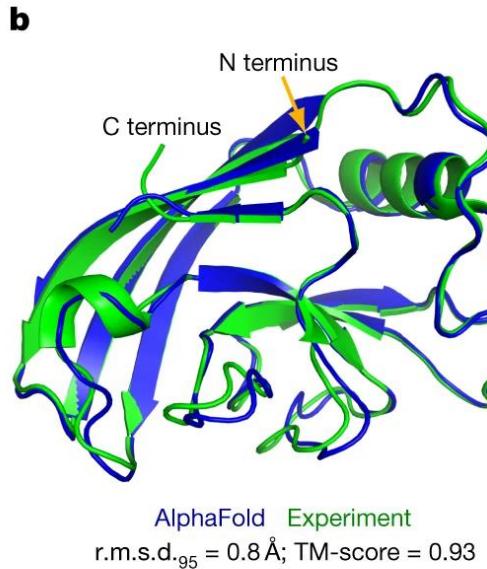
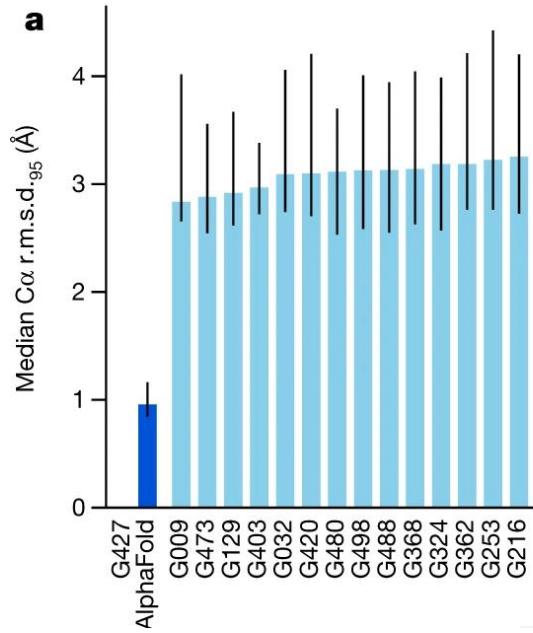
- *Ab Initio* or “*de novo*”

- When you have no homologous/fold template
- Physics based methods
 - E.g. Molecular Dynamics – Anton, D.E. Shaw; Folding@home
- Fragment based methods
 - E.g. Rosetta
- **Computationally Expensive!**



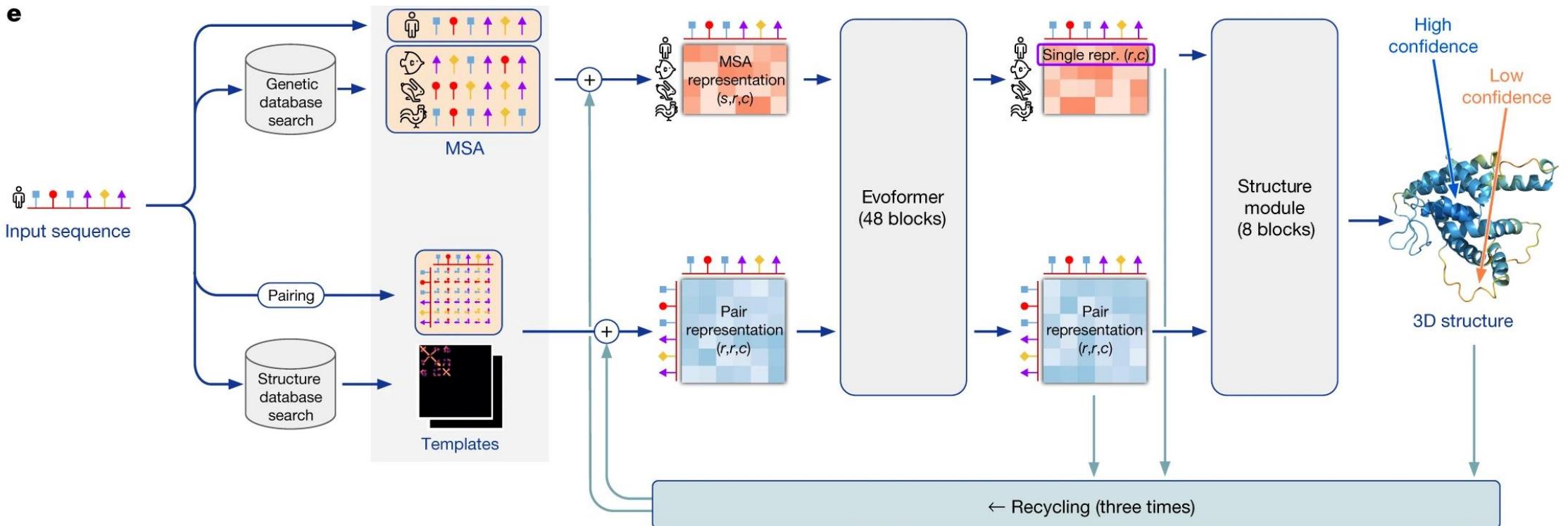
Cutting-Edge

- AlphaFold2
- Deepmind



Cutting-Edge

- “End-to-End differentiable”
- Can use templates
- New version can predict complexes!



Cutting-Edge

- **AlphaFold2**
- Deepmind
- Database in collaboration with the EBI

AlphaFold Protein Structure Database

Home About FAQs Downloads

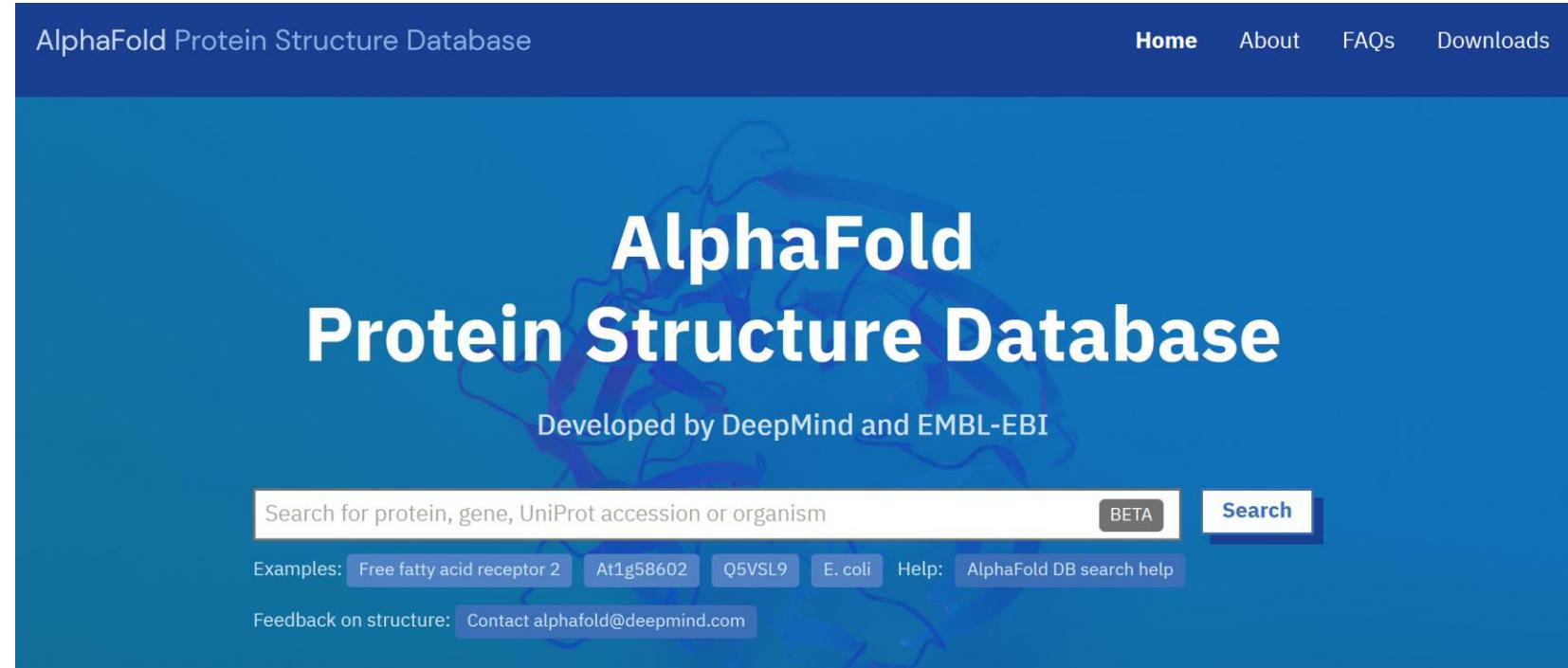
AlphaFold Protein Structure Database

Developed by DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism BETA Search

Examples: Free fatty acid receptor 2 At1g58602 Q5VSL9 E. coli Help: AlphaFold DB search help

Feedback on structure: Contact alphafold@deepmind.com



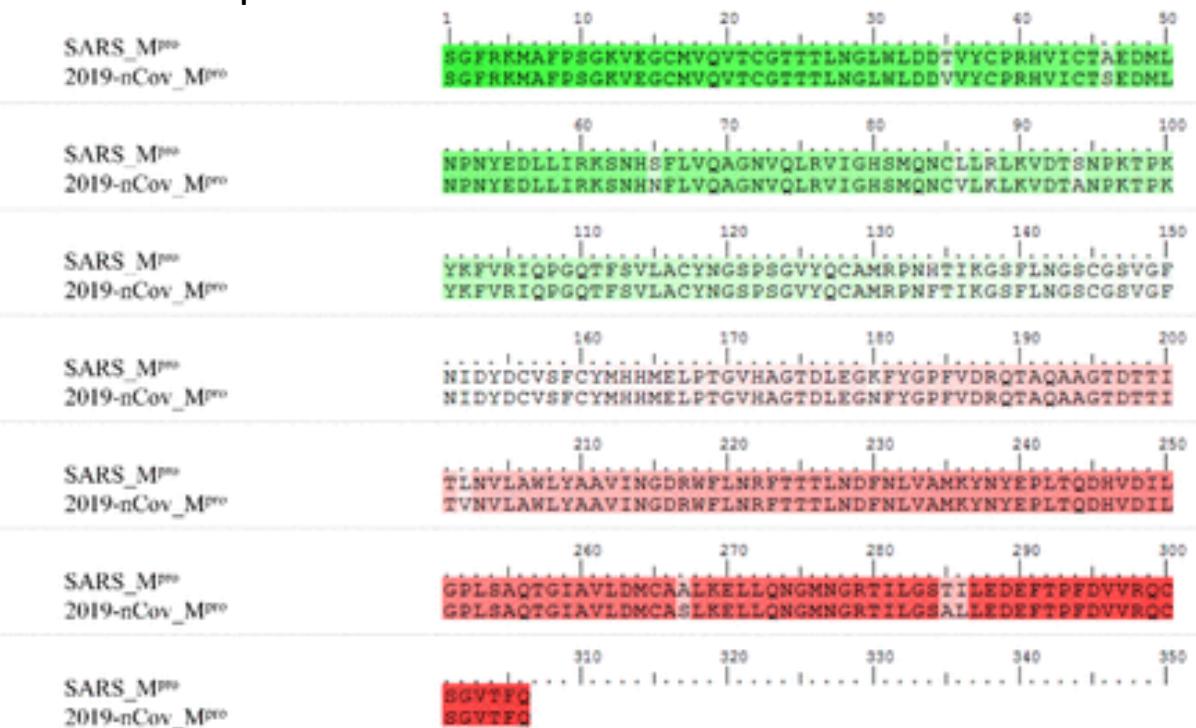
<https://alphafold.ebi.ac.uk/>

Case study

SARS-CoV-2 main protease Xu et al. January 2020

- Try and identify drug treatments

High Sequence identity with SARS-CoV-1 proteins



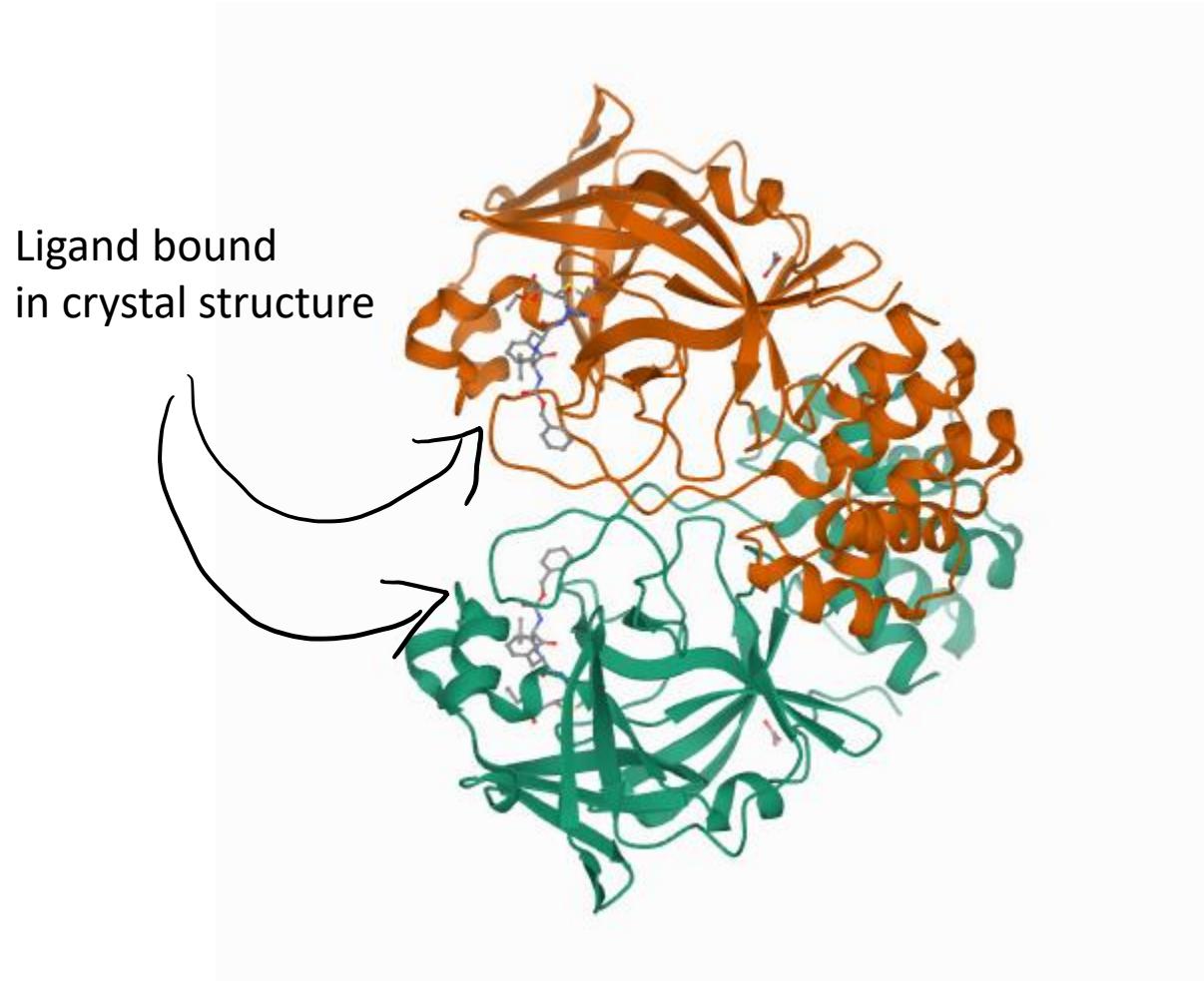
Case study

2GTB - template

SARS-CoV-2 main protease

Xu et al. January 2020

- Try and identify drug treatments
- Homology modelling of main protease
 - SWISS-MODEL

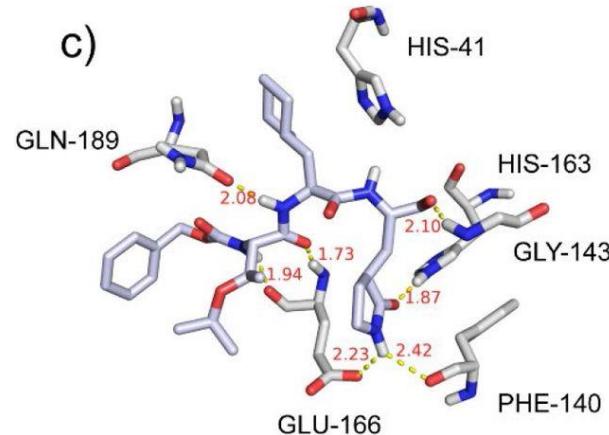
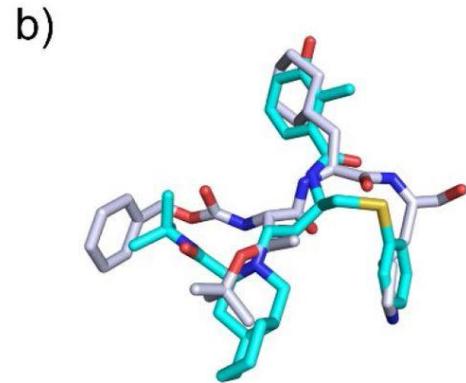
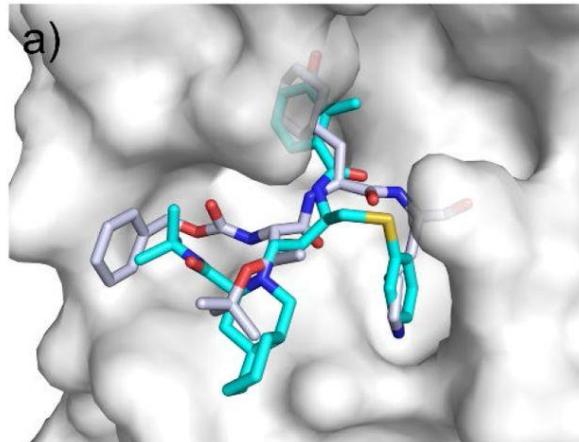


Case study

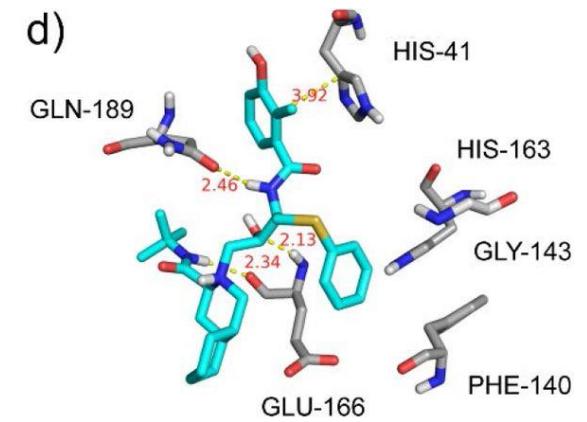
SARS-CoV-2 main protease
Xu et al. January 2020

- Try and identify drug treatments
- Homology modelling of main protease
 - SWISS-MODEL
- Virtual screening of ~2000 drugs to the models
 - Autodock Vina

Nelfinavir



Original Ligand



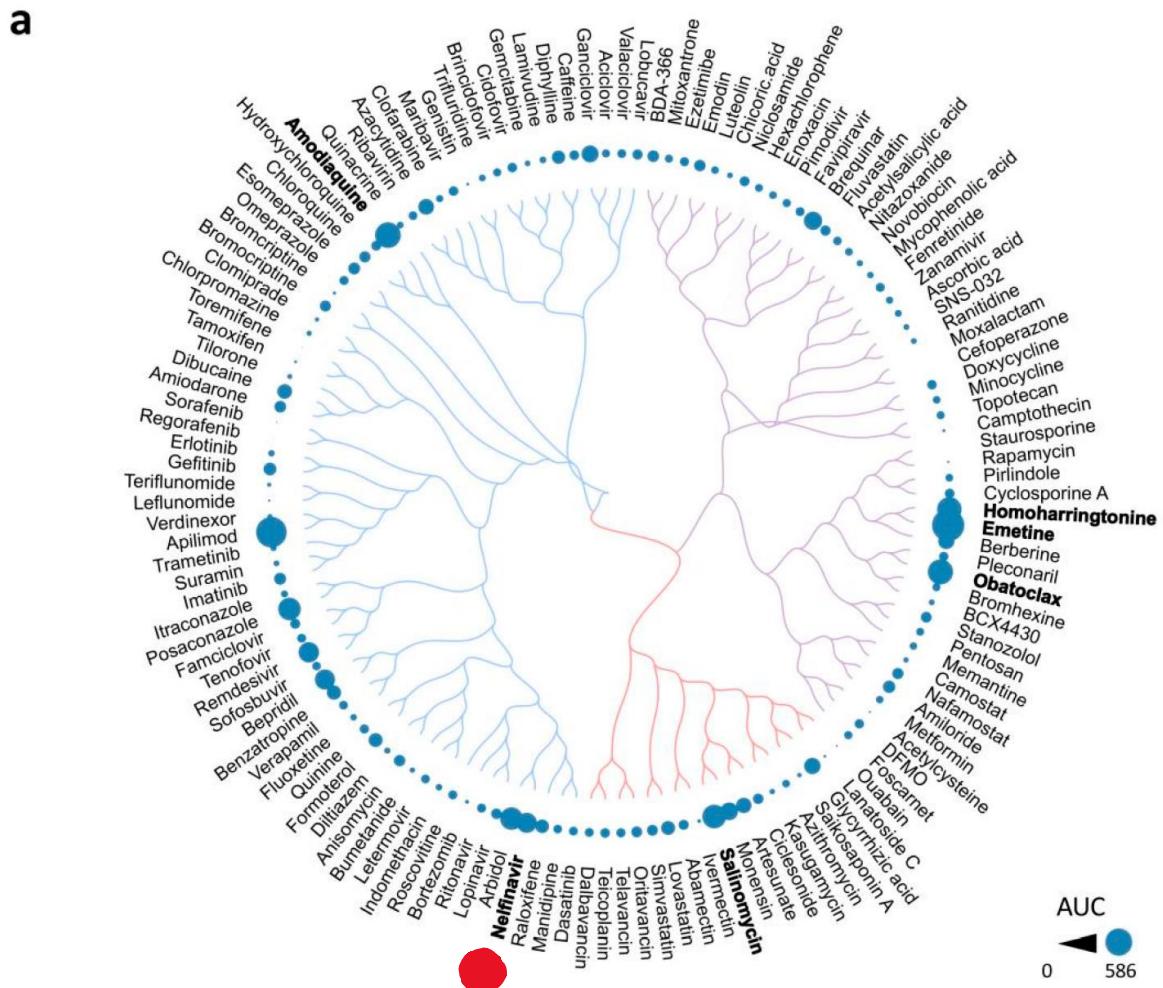
Bound Drug

Case study

SARS-CoV-2 main protease

Ianevski et al. June 2020

- Subsequent in vitro studies showed Nelfinavir inhibits viral infection in cell cultures
 - As of earlier this year (2021), clinical trials in patients



Summary

- Structural bioinformatics is the analysis and prediction of the 3D structures of biological macromolecules
- We can predict both secondary and tertiary structure from sequence
 - Central dogma – sequence -> structure -> function
 - Evolutionary information key!
- Structure is more conserved than sequence
 - Fold space is limited
- Template-based modelling vs Template-Free modelling
- Cutting edge is Deep Learning – AlphaFold 2
- Modelling combined with function prediction e.g. docking, molecular dynamics allows for fast hypothesis generation *in silico*