NEAR EAST UNIVERSITY
DESAM INSTITUTE

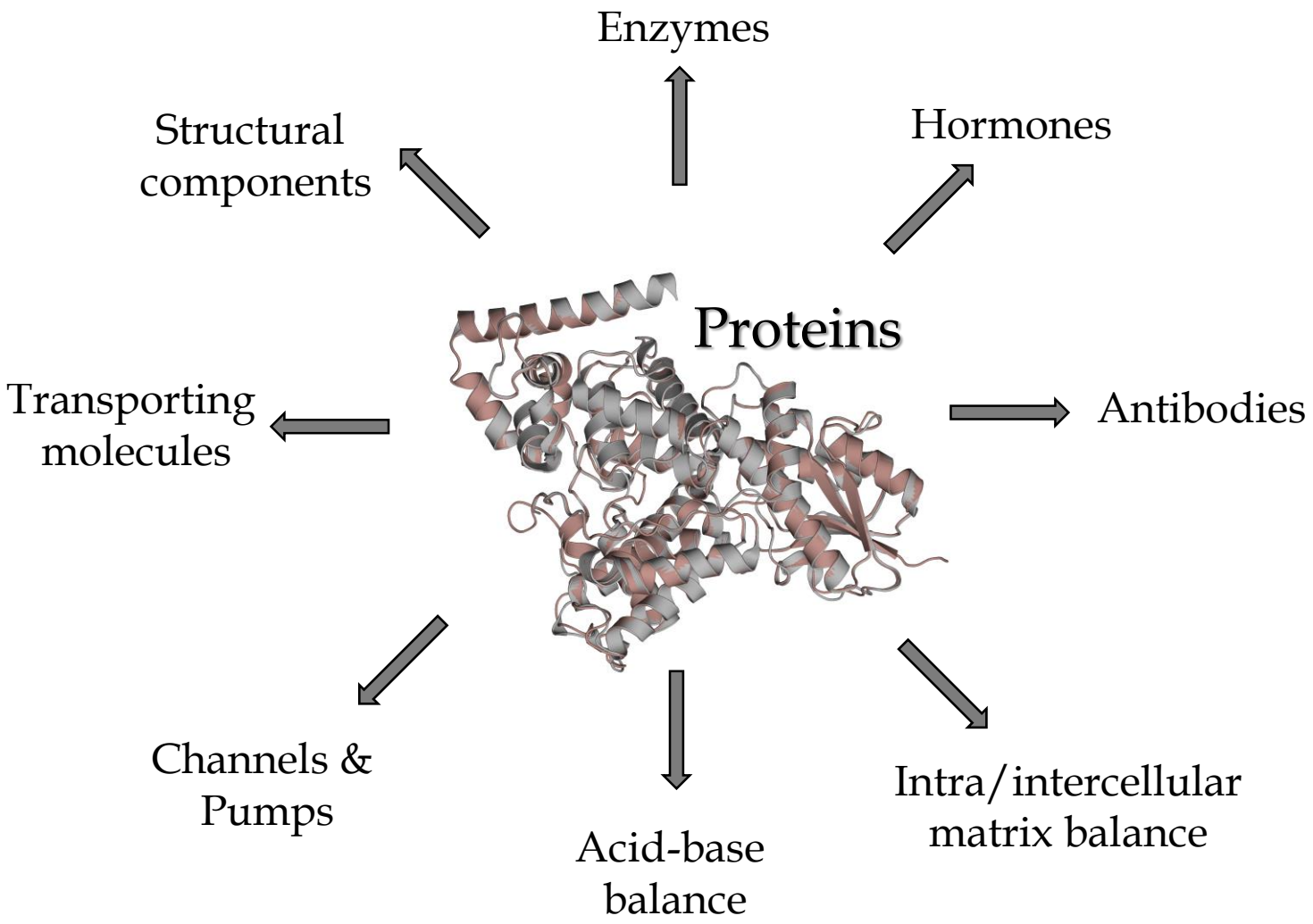
II. BIOINFORMATICS WINTER SCHOOL:
COMPUTER METHODS IN MOLECULAR SCIENCES

In Silico Homology Modelling of Proteins

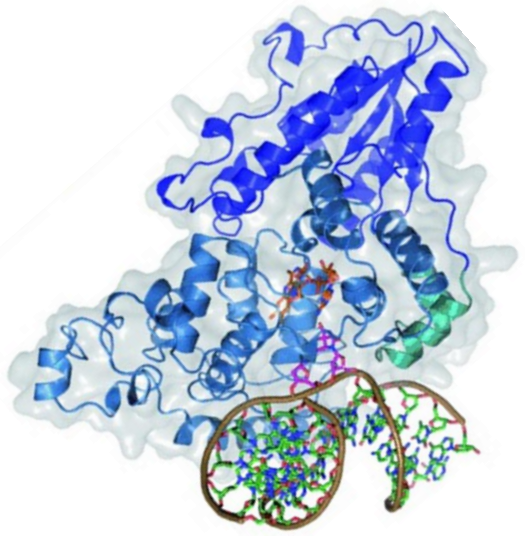
Şeref Gül, PhD.
Chemical and Biological Engineering
Koç University

KOÇ UNIVERSITY

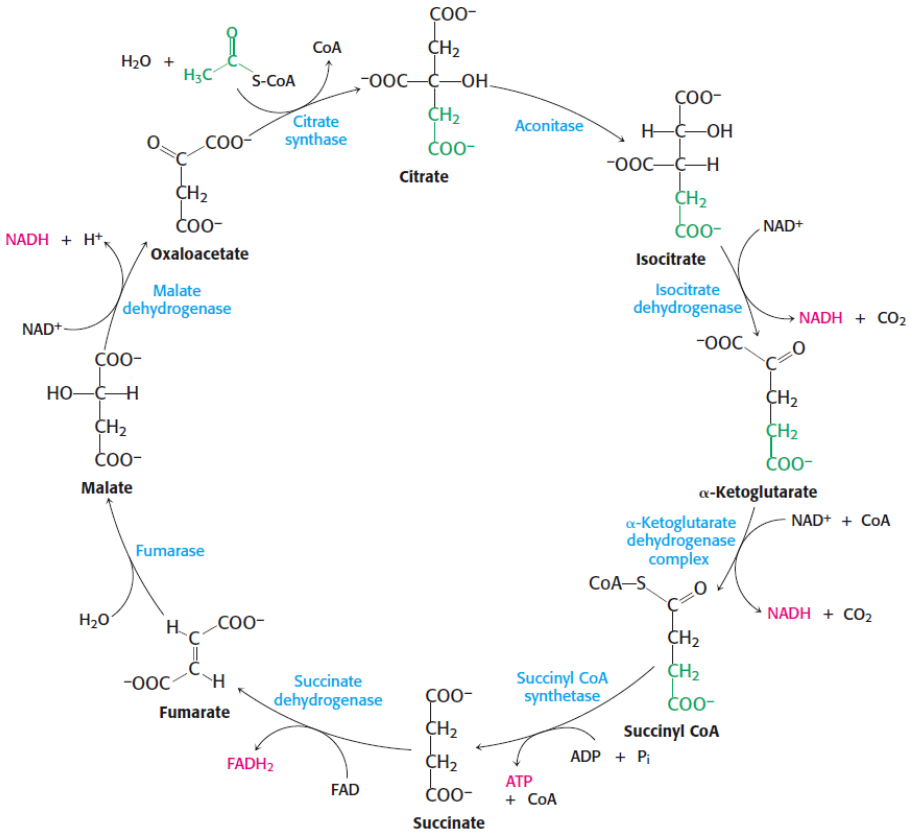
Roles of Proteins

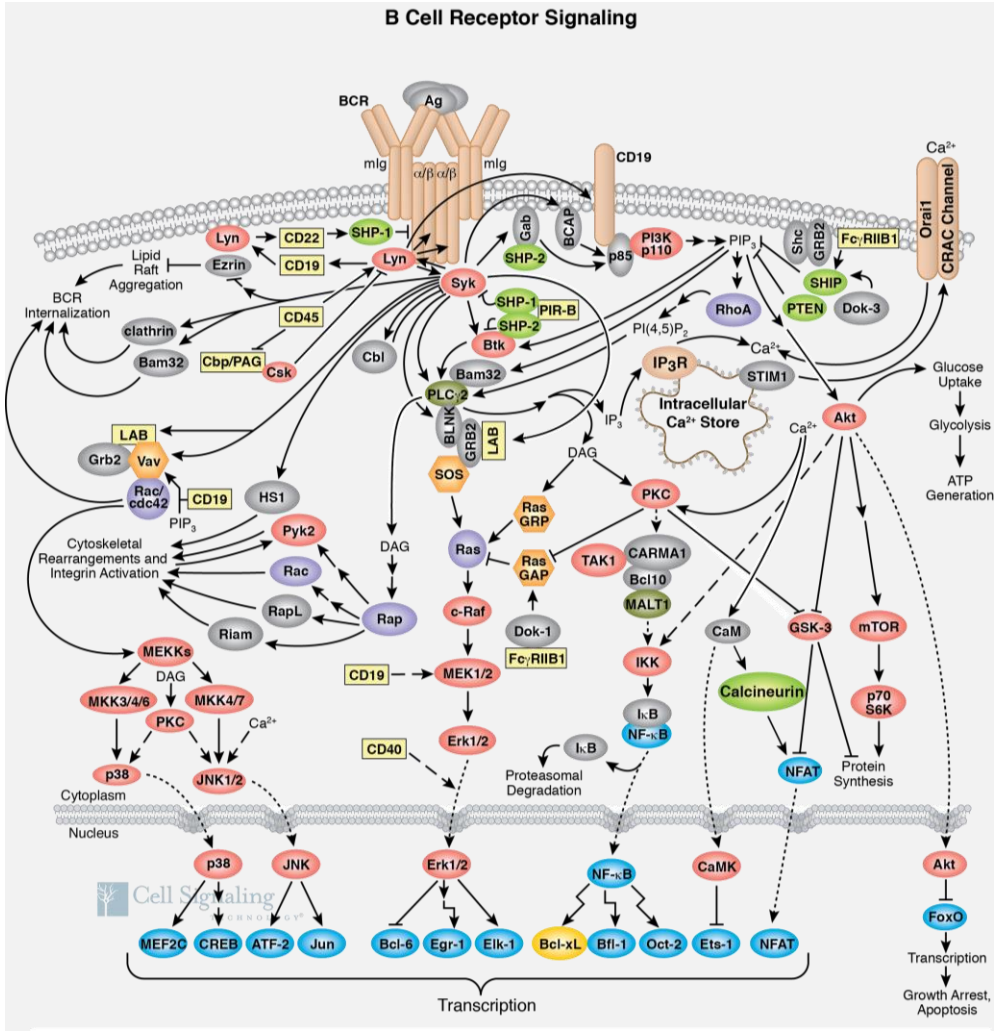


Protein-Protein Interactions are Vital for Life



Angew. Chem. Int. Ed., **2008**, *47*, 10.1002/anie.200804268





Structure Information of Proteins

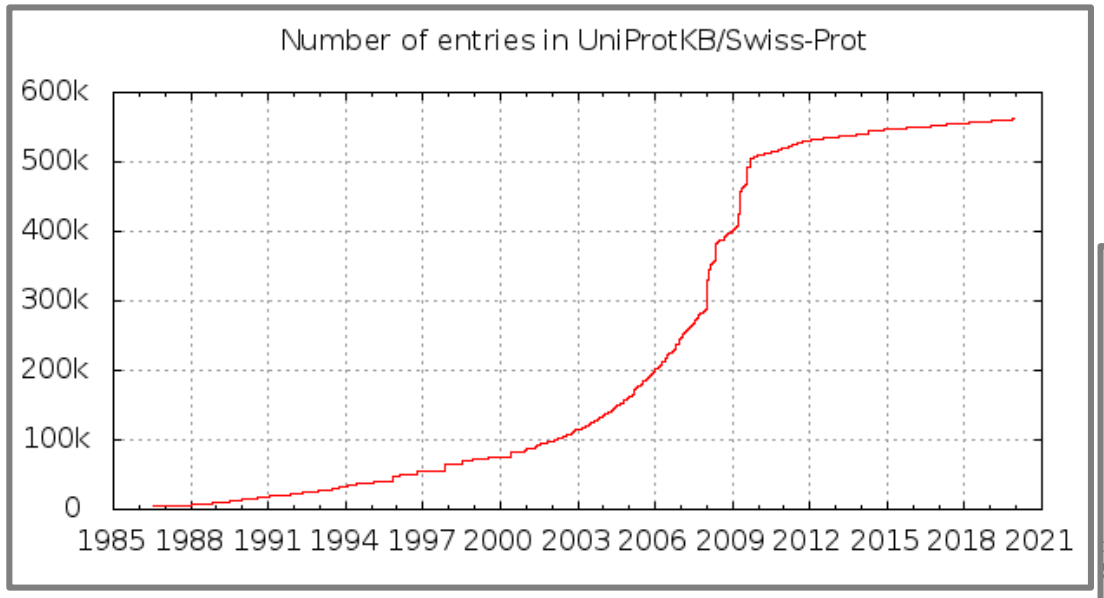
- Enhances our understanding of how proteins interaction with each other
  - How they function; elucidating pathways in the cell

- Useful to understand mechanism of diseases, e.g. Alzheimer's, Parkinson's, cystic fibrosis, and Hungtinton's, as a result of misfolded proteins

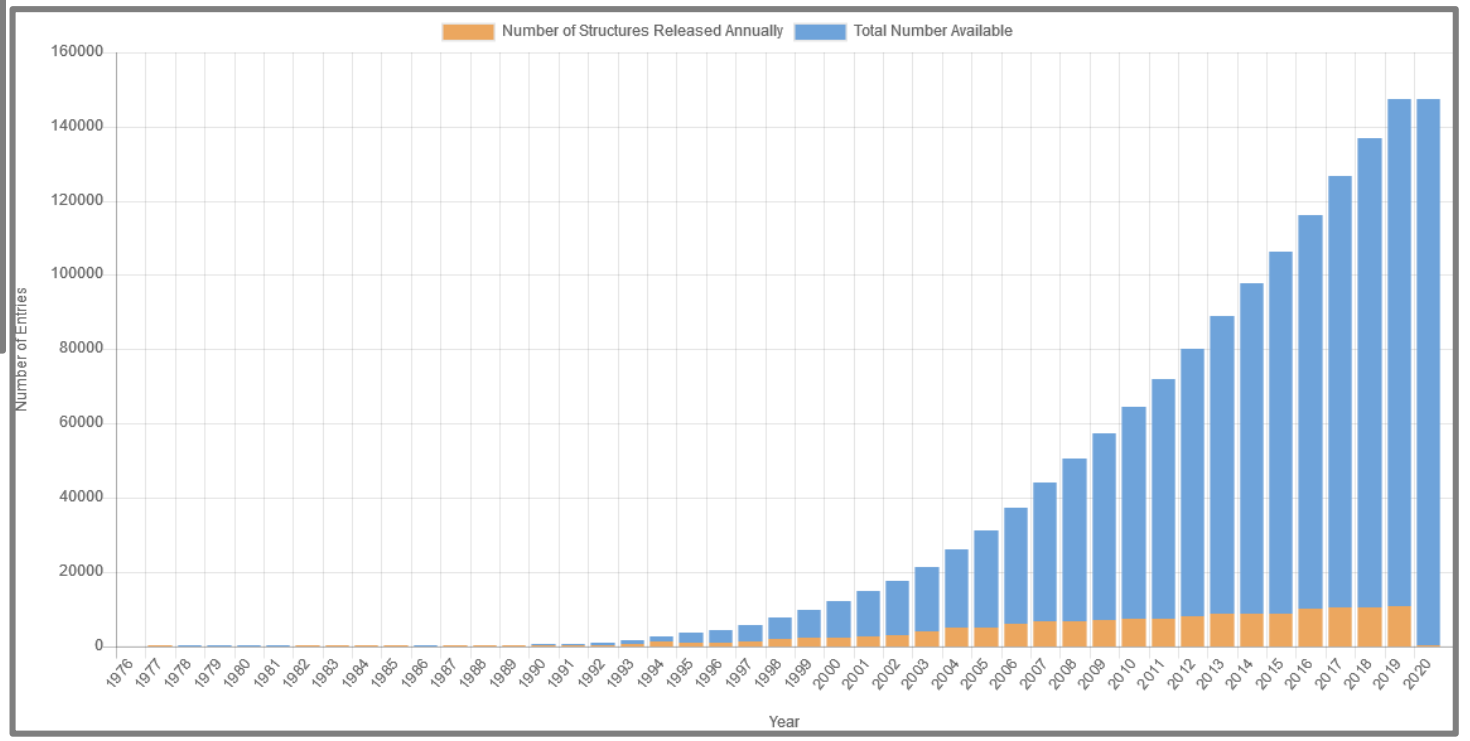
- Can be utilized in drug design

- Helps to design site-directed mutagenesis

UniprotKB/Swiss-Prot vs PDB

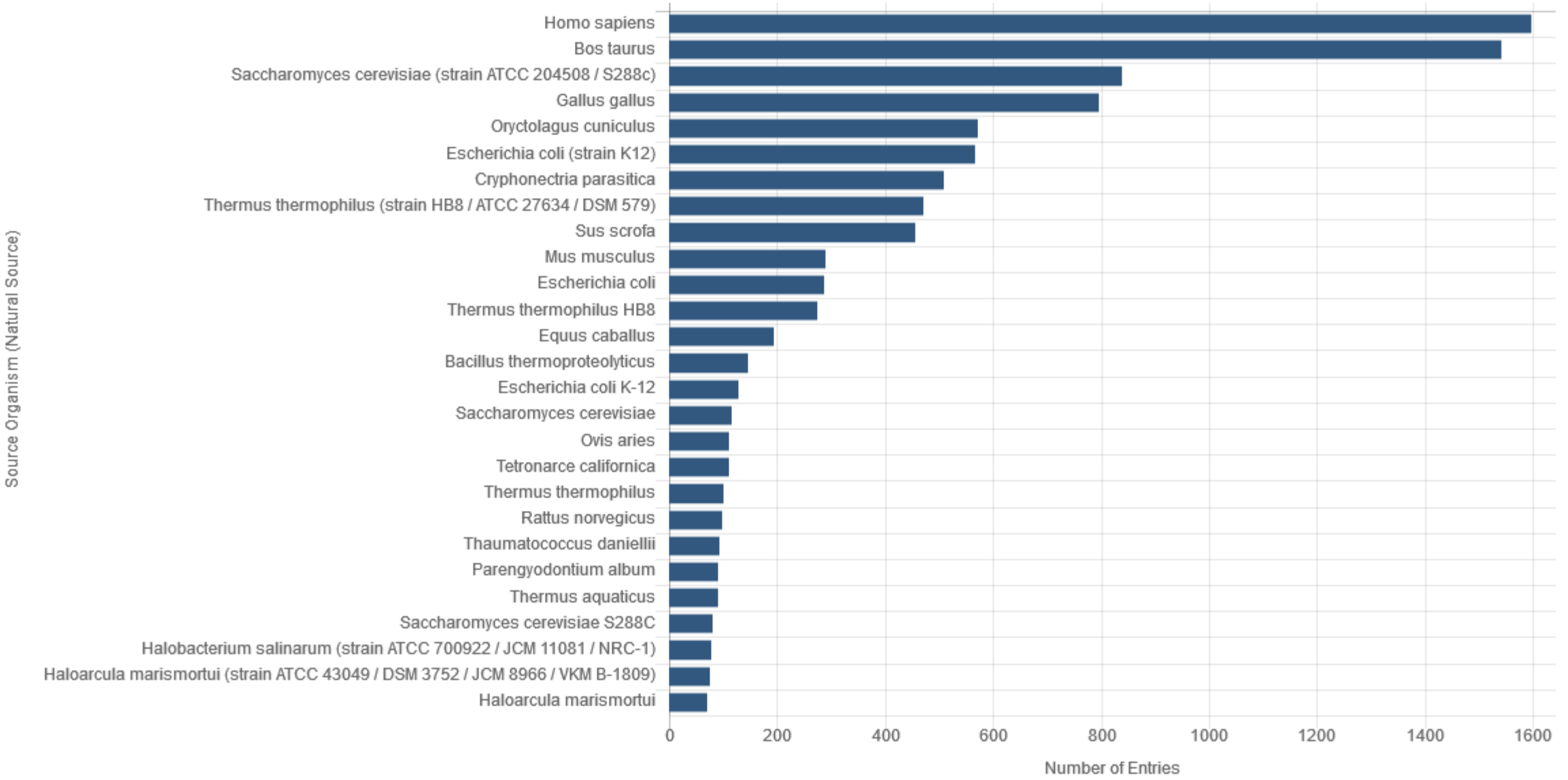


https://web.expasy.org/docs/relnotes/relstat.html



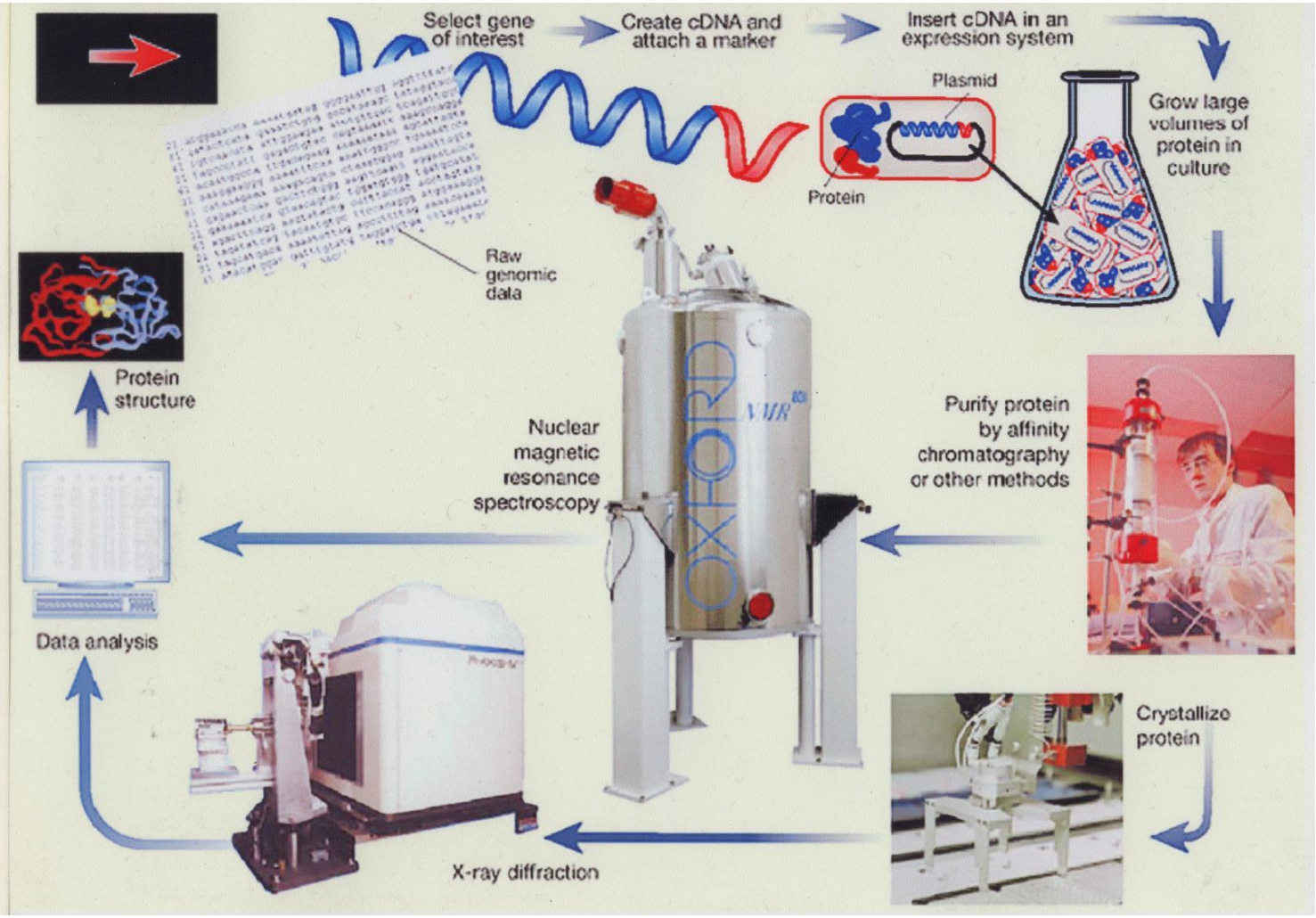
- Yearly ~10000 protein structure are resolved

Protein Data Bank



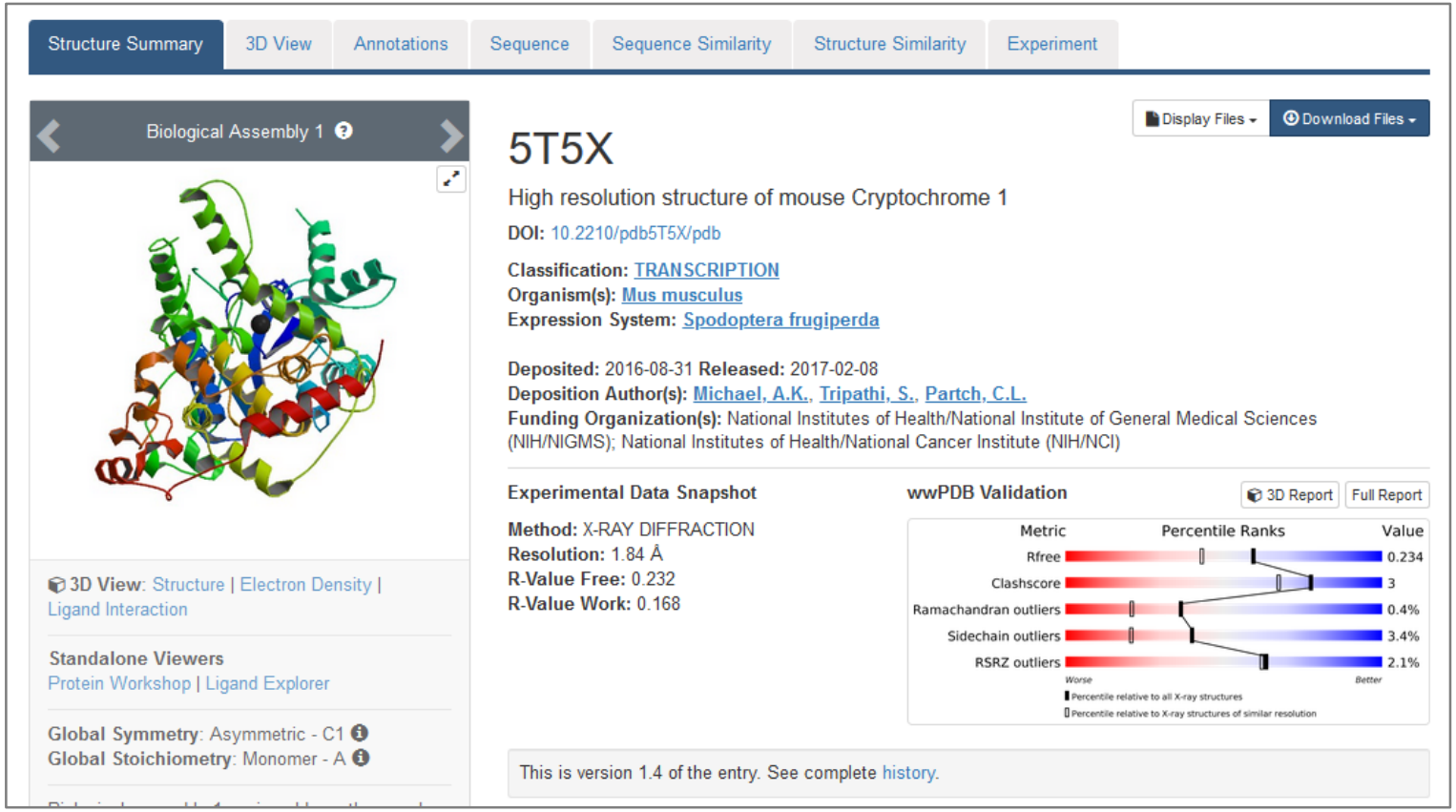
- Number of proteins predicted in human: ~20000 if 1 gene = 1 protein; ~75000 protein entries in UNIPROT.
- Alternative splicing, SNPs may produce ~100 protein from a protein

Experimental Determination of Protein Structure



- Data collection is slow and laborious
- Some methods require expertise
- Some proteins fail to crystallize
- Large amount of protein requirements for NMR can be problematic
- Size of protein is limiting factor for NMR

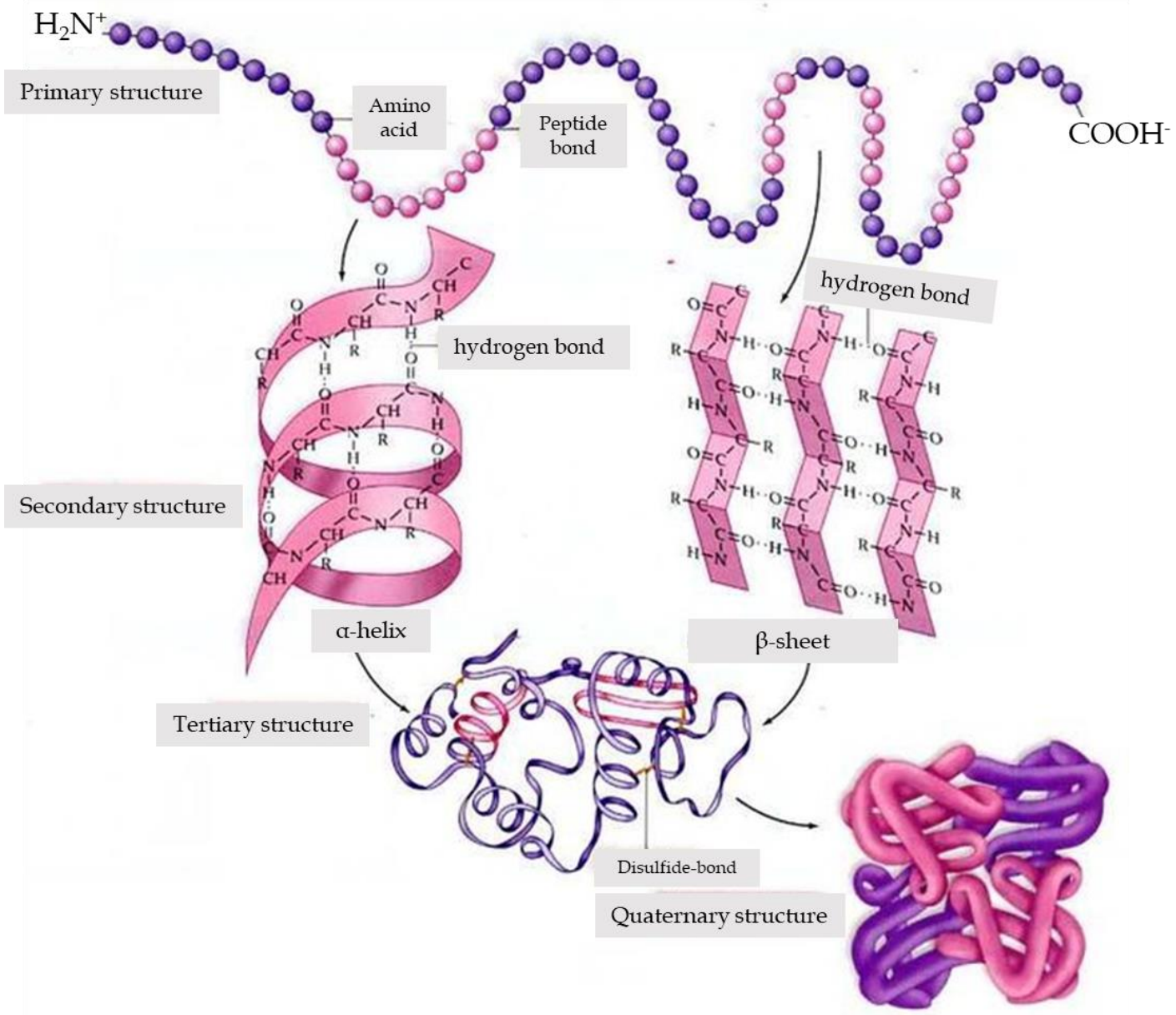
Protein Info in PDB



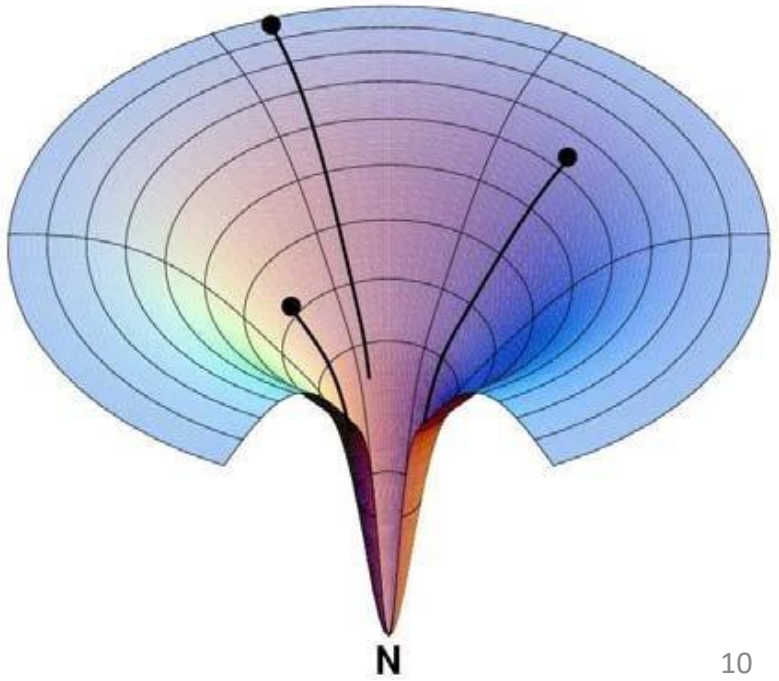

```
HEADER    TRANSCRIPTION                           31-AUG-16
5T5X
TITLE     HIGH RESOLUTION STRUCTURE OF MOUSE CRYPTOCHROME 1
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: CRYPTOCHROME-1;
COMPND   3 CHAIN: A;
COMPND   4 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: MUS MUSCULUS;
SOURCE   3 ORGANISM_COMMON: MOUSE;
SOURCE   4 ORGANISM_TAXID: 10090;
SOURCE   5 GENE: CRY1;
SOURCE   6 EXPRESSION_SYSTEM: SPODOPTERA FRUGIPERDA;
SOURCE   7 EXPRESSION_SYSTEM_TAXID: 7108
KEYWDS    REPRESSOR, TRANSCRIPTION, CIRCADIAN
EXPDTA    X-RAY DIFFRACTION
AUTHOR    A.K.MICHAEL,S.TRIPATHI,C.L.PARTCH
REVDAT   3   22-FEB-17 5T5X    1       JRNL
REVDAT   2   15-FEB-17 5T5X    1       JRNL
REVDAT   1   08-FEB-17 5T5X    0
JRNL        AUTH
A.K.MICHAEL,J.L.FRIBOURGH,Y.CHELLIAH,C.R.SANDATE,G.L.HURA,
JRNL        AUTH 2 D.SCHNEIDMAN-
DUHOVNY,S.M.TRIPATHI,J.S.TAKAHASHI,C.L.PARTCH
JRNL        TITL   FORMATION OF A REPRESSIVE COMPLEX IN THE
MAMMALIAN CIRCADIAN
JRNL        TITL 2 CLOCK IS MEDIATED BY THE SECONDARY POCKET OF
CRY1.
JRNL        REF    PROC. NATL. ACAD. SCI.        V. 114   1560
2017
JRNL        REF  2 U.S.A.
JRNL        REFN                   ESSN 1091-6490
JRNL        PMID   28143926
JRNL        DOI    10.1073/PNAS.1615310114
REMARK   2
REMARK   2 RESOLUTION.    1.84 ANGSTROMS.
```

We need to predict the structure of proteins

What Determines the Structure of a Protein



- Assume an amino acid can adopt 10 conformations
- For a peptide w 100 residues => 10^{100} conformations
- If a conformation is analyzed in 10^{-13}sec
- ~10^{68} years to sample all conformations
- In cells proteins fold in ms time scale



Adopted from S. Mader, *Biology*. McGraw Hill, 2010

Protein Structure Prediction

De novo methods

Ab initio

- MD method
- Search for global minimum free energy

- No need for template
- Allows to discover new folds

- But limited to local minima
- Can be used for small proteins <120aa

Ab initio w database info

- Utilizes folding pattern of similar sub-sequences of target
- Assembled into a low E structure via optimization

- Conformational search space is limited compared to *ab initio* method

- Protein may obtain many conformation

Template based methods

Homology/comparative modelling

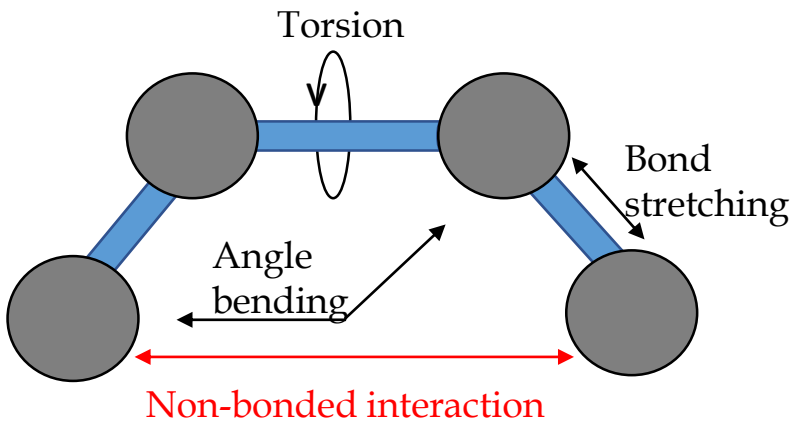
- Depend on available structure(s)
- Provide more accurate protein structure models

- Able to predict structure of longer protein

- May benefit from multiple templates

- Restricted to available structures

De Novo Modelling



$$U(R) = \sum_{bonds} k_r \left(r - r_{eq}\right)^2$$ *bond*

$$+ \sum_{angles} k_\theta \left(\theta - \theta_{eq}\right)^2$$ *angle*

$$+ \sum_{dihedrals} k_\phi \left(1 + \cos\left[n\phi - \gamma\right]\right)$$ *dihedral*

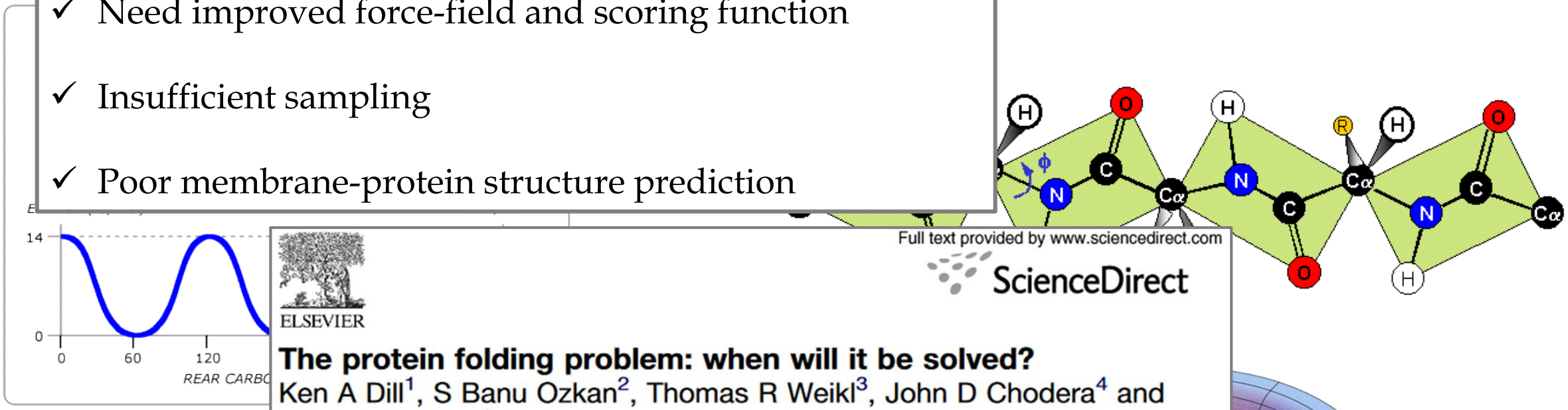
$$+ \sum_{impropers} k_\omega \left(\omega - \omega_{eq}\right)^2$$ *improper*

$$+ \sum_{i<j}^{atoms} \varepsilon_{ij} \left[\left(\frac{r_m}{r_{ij}} \right)^{12} - 2 \left(\frac{r_m}{r_{ij}} \right)^{6} \right]$$ *Van der Waals*

$$+ \sum_{i<j}^{atoms} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$$ *electrostatic*

De Novo Modelling

- ✓ Need improved force-field and scoring function
- ✓ Insufficient sampling
- ✓ Poor membrane-protein structure prediction



ELSEVIER

The protein folding problem: when will it be solved?

Ken A Dill[1], S Banu Ozkan[2], Thomas R Weikl[3], John D Chodera[4] and Vincent A Voelz[4]

Current Opinion in Structural Biology 2007, **17**:342–346

- ❖ Assume an amino acid can adopt 10 con
- ❖ For a peptide w 100 residues => 10^{100} co
- ❖ If a conformation is analyzed in 10^{-13}sec
- ❖ ~10^{68} years to sample all conformations
- ❖ In cells proteins fold in ms time scale

Advanced Review

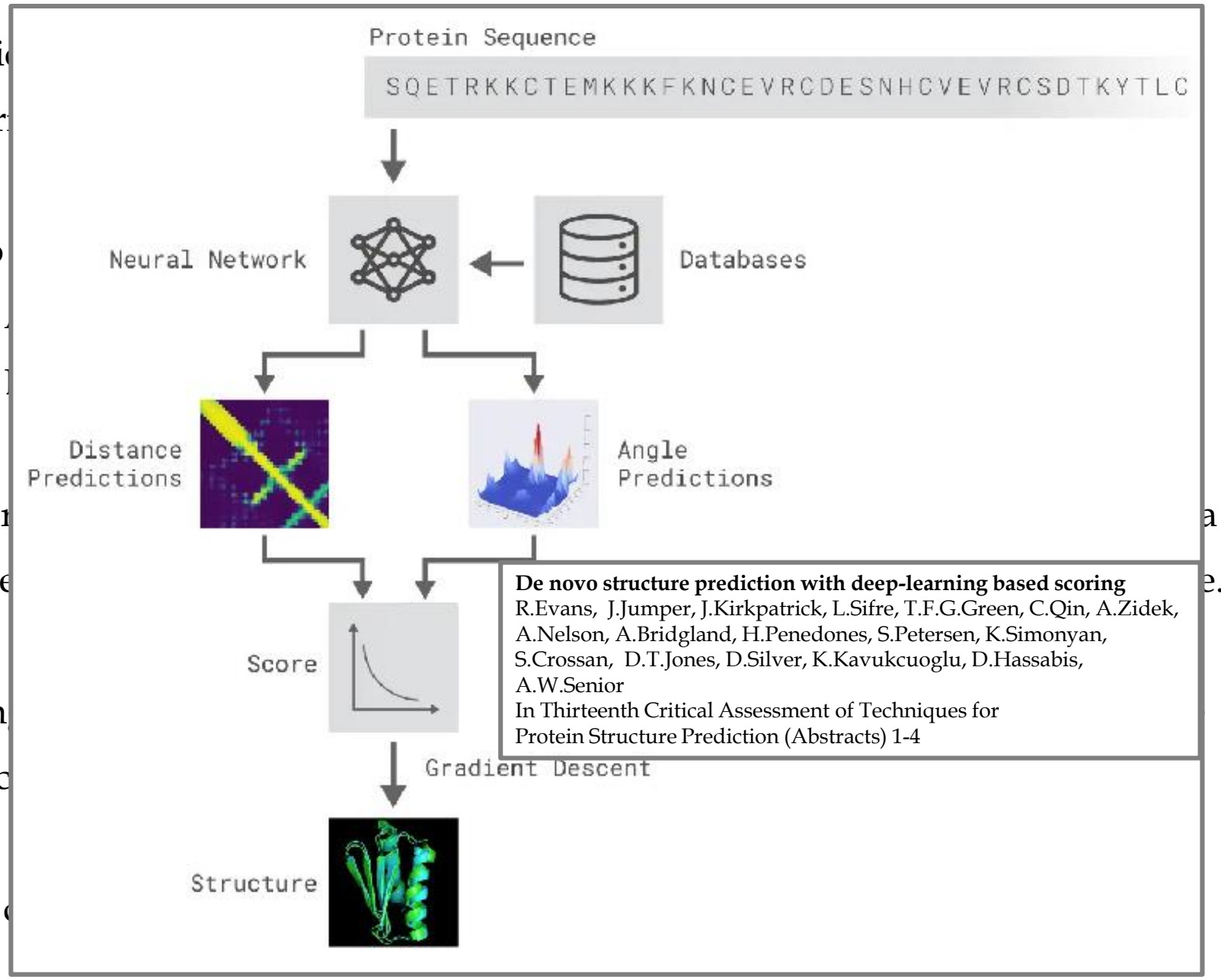
Computational protein structure refinement: almost there, yet still so far to go

Michael Feig*

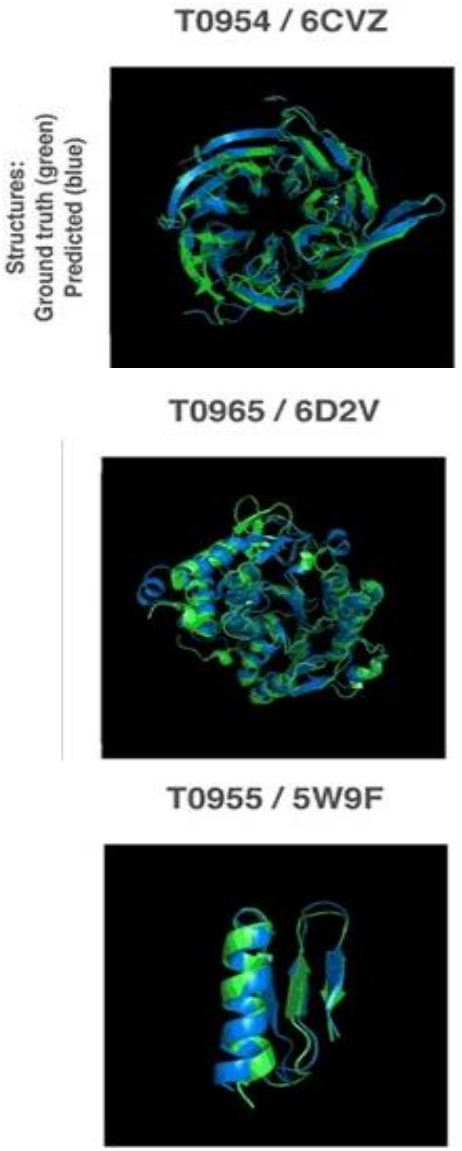
WIREs Comput Mol Sci 2017, 7:e1307. doi: 10.1002/wcms.1307

De Novo Modelling – Artificial Intelligence

- ✓ Criti
  inter
- ✓ Two
  - ✓
  - ✓
- ✓ Neur
  prote
- ✓ Usin
  struc
- ✓ Full



De novo structure prediction with deep-learning based scoring
R.Evans, J.Jumper, J.Kirkpatrick, L.Sifre, T.F.G.Green, C.Qin, A.Zidek, A.Nelson, A.Bridgland, H.Penedones, S.Petersen, K.Simonyan, S.Crossan, D.T.Jones, D.Silver, K.Kavukcuoglu, D.Hassabis, A.W.Senior
In Thirteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstracts) 1-4



Homology Modelling

- Homology modelling = Comparative modelling
- Originates from "similar sequences exhibit similar structures"
- Resolved structure (w similar 1^0 structure) is used as the **template** to model the protein
- Usage/Advantages
  - May generate structure which can not be obtained via experimental
  - May provide structural information of a mutant protein causing a disease
  - Can be utilized in structure-based drug design studies

Steps of Homology Modelling

- ✓ Template recognition / finding homologous sequences in PDB
- ✓ Sequence alignment
- ✓ Determining structurally conserved regions
- ✓ Determining structurally variable regions
- ✓ Backbone generation depending on conserved regions
- ✓ Conformational search for side chains
- ✓ Refinement of structure
- ✓ Validation of model

Template Recognition / Sequence Alignment

- Known protein structure having high homology can be used to model the homolog unknown protein

Query Sequence

```
MRYSVVRLILGDQLNHAHSWFSEHRDDVLYLIAELH
QEQEYVRHHIQKQCAFFAAMQAFADYLSAEGHHV
WHLDLDASAQYNDLPDLIAQICQQVQADAFQYQRP
DEYRLLEQMANLRLSGITIGCVDTEHFLLPFAEIP
EQFPASKAVLMEHFYRRMRKRFGYLMTADGKPEGG
QWNFDADNRNKLKSPDLLQLPTPLCFDNPVASIKA
RIERHRIPSIGQVGESLLWPINRAQALSLLAHFCQICL
PNFGRFQDAMTAQHPHRWSLYHSRLSFALNSK
```

#1

```
LVLGDQLSDDLPALRAADPAADLVVMAEVMEEGTY
VPHHPQKIALILAAMRKFARRLQERGFRVAYSRLD
DPDTGPSIGAELLRRAAETGAREAVATRPGDWRLIEA
LEAMPLPVRFLPDDRFLCPADEFARWTEGRKQL
RMEWFYREMRRRTGLLMEGDEPAGGKWNFDTENR
KPAAPDLLRPRPLRFEPDAEVRAVLDLVEARFPRHF
GRLRPFHWATDRAEALRALDHFIRESLPRFGDEQDA
MLADDPFLSHALLSSSMNLGLLGPMEVCRRAETE
```

#2

```
MSQLVLILGDQLSPSIAALDGVDKKQDTIVLCEVMAE
ASYVGHHKKKIAFWFSAMRHFAEELRGEGYRVR
YTRIDDADNAGSFTGEVKRAIDDLTPSRICVTEPGEW
RVRSEMDGFAGAFGIQVDIRSDRRFLSSHGEFR
NWAAGRKSLTMEYFYREMRRKTGLLMNGEQPVGG
RWNFDAENRQPARPDLLRPKHPVFAPDKITKEVIDT
VERLFPDNFGKLENFGFAVTRTDAERALSAFIDDFLC
NFGATQDAMLQDDPNLNHSLLSFYINCGLLDAL
```

#3

```
MSQLVLILGDQLSPSIAALDGVDKKQDTIVLCEVMAE
ASYVGHHKKKIAFIFSAMRHFAEELRGEGYRVR
YTRIDDADNAGSFTGEVKRAIDDLTPSRICVTEPGEW
RVRSEMDGFAGAFGIQVDIRSDRRFLSSHGEFR
NWAAGRKSLTMEYFYREMRRKTGLLMNGEQPVGG
RWNFDAENRQPARPDLLRPKHPVFAPDKITKEVIDT
VERLFPDNFGKLENFGFAVTRTDAERALSAFIDDFLC
NFGATQDAMLQDDPNLNHSLLSFYINCGLLDAL
```

#4

```
MSQLVLILGDQLSPSIAALDGVDKKQDTIVLCEVMAE
ASYVGHHKKKIAFIFSAMRHFAEELRGEGYRVR
YTRIDDADNAGSFTGEVKRAIDDLTPSRICVTEPGEW
RVRSEMDGFAGAFGIQVDIRSDRRFLSSHGEFR
NWAAGRKSLTMEYFYREMRRKTGLLMNGEQPVGG
RWNFDAENRQPARPDLLRPKHPVFAPDKITKEVIDT
VERLFPDNFGKLENFGFAVTRTDAERALSAFIDDFLC
NFGATQDAMLQDDPNLNHSLLSFYINCGLLDAL
```

Template Recognition / Sequence Alignment

	G	E	N	E	T	I	C	S
G	10	0	0	0	0	0	0	0
E	0	10	0	10	0	0	0	0
N	0	0	10	0	0	0	0	0
E	0	0	0	10	0	0	0	0
S	0	0	0	0	0	0	0	10
I	0	0	0	0	0	10	0	0
S	0	0	0	0	0	0	0	10

	G	E	N	E	T	I	C	S
G	60	40	30	20	20	0	10	0
E	40	50	30	30	20	0	10	0
N	30	30	40	20	20	0	10	0
E	20	20	20	30	20	10	10	0
S	20	20	20	20	20	0	10	10
I	10	10	10	10	10	20	10	0
S	0	0	0	0	0	0	0	10

Sequence 1: G-E-N-E-T-I-C-S
Sequence 2: G-E-N-E-S-I- - -S

from David Wishart

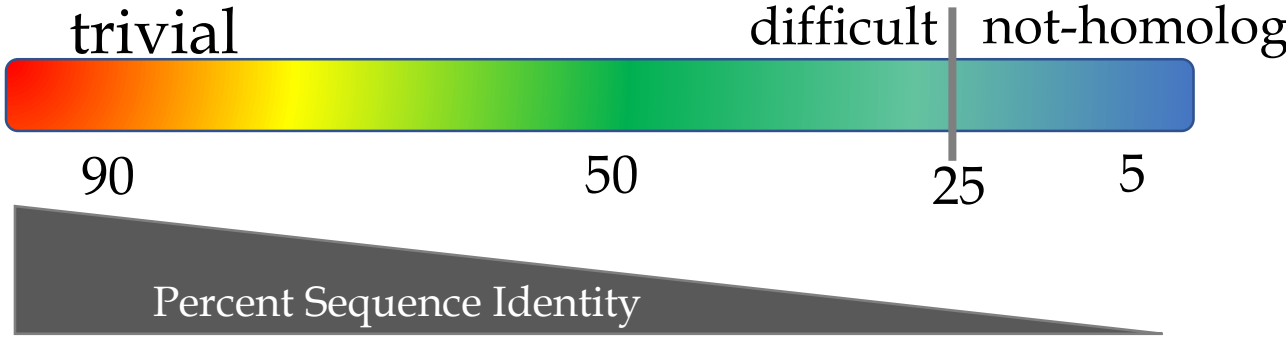
- e.g. PAM and BLOSUM scoring matrices
- Global vs local alignment
- Minor errors in alignment can cause major differences in model structures

Template Recognition / Sequence Alignment

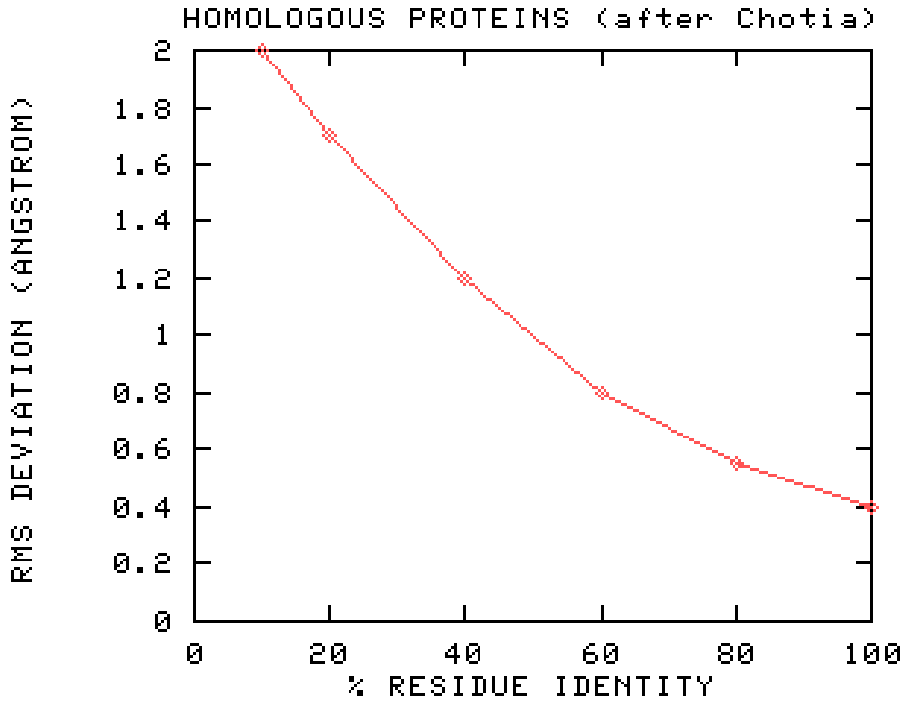
- Known protein structure having high homology can be used to model the homolog unknown protein

Query Sequence

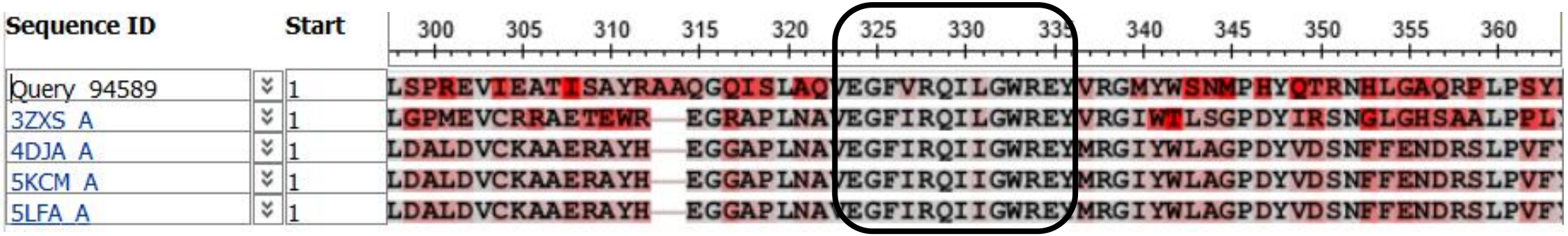
```
MRYSVVRLILGDQLNHAHSWFSEHRDDVLYLIAELHQE
QEYVRHHIQKQCAFFAAMQAFADYLSAEGHHV
WHLDLDASAQYNDLPDLIAQICQQVQADAFQYQRPDE
YRLLEQMANLRLSGITIGCVDTEHFLLPFAEIP
EQFPASKAVLMEHFYRRMRKRFGYLMTADGKPEGGQ
WNFDADNRNKLKSPDLLQLPTPLCFDNPVASIKA
RIERHRIPSIGQVGESLLWPINRAQALSLLAHFCQICLPN
FGRFQDAMTAQHPHRWSLYHSRLSFALNSK
```




```
Method:Compositional matrix adjust.,
Identities:217/511(42%), Positives:296/511(57%), Gaps:15/511(2%)

Query  8    LILGDQL-NHAHSWFSEHRDDVLYLIAELHQEQEYVRHHIQKQCAFFAAMQAFADYLSAE  66
            L+LGDQL +    +  +       L ++AE+ +E   YV HH QK      AAM+ FA  L
Sbjct  20   LVLGDQLSDDLPALRAADPAADLVVMAEVMEEGTYVPHHPQKIALILAAMRKFARRLQER  79

Query  67   GHHVWHLDLDASAQYNDLPDLIAQICQQVQADAFQYQRPDEYRLLEQMANLRLSGITIGC  126
            G  V +  LD        +    + +    +  A       RP ++RL+E +  + L    +
Sbjct  80   GFRVAYSRLDDPDTGPSIGAELLRRAAETGAREAVATRPGDWRLIEALEAMPLP---VRF  136

Query  127  VDTEHFLLPFAEIPEQFPASKAVLMEHFYRRMRKRFGYLMTADGKPEGGQWNFDADNRNK  186
            +   + FL P   E         K + ME FYR MR+R G LM   D +P GG+WNFD +NR K
Sbjct  137  LPDDRFLCPADEFARWTEGRKQLRMEWFYREMRRRTGLLMEGD-EPAGGKWNFDTENR-K  194

Query  187  LKSPDLLQLPTPLCFDNPVASIKARIE--RHRIPSIGQVGESLLWPINRAQALSLLAHFC  244
              +PDLL+ P PL F+ P A ++A ++    R P          W  +RA+AL  L HF
Sbjct  195  PAAPDLLR-PRPLRFE-PDAEVRAVLDLVEARFPRHFGRLRPFHWATDRAEALRALDHFI  252
```

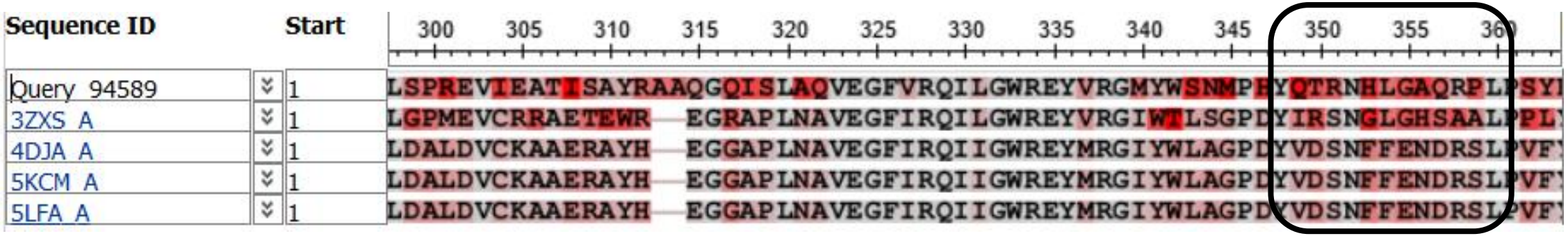


Finding Structurally Conserved Regions



- Regions correspond to least number of gap & high level of conservation
- Conserved regions refer to stable region of protein; generally located inside the protein
- Generally corresponds to secondary structure

Finding Structurally Variable Regions



- Regions correspond to the greatest number of gap & least level of conservation
- Variable regions refer to least stable region or most flexible region of protein; generally located outside the protein
- Generally corresponds to loops and turns

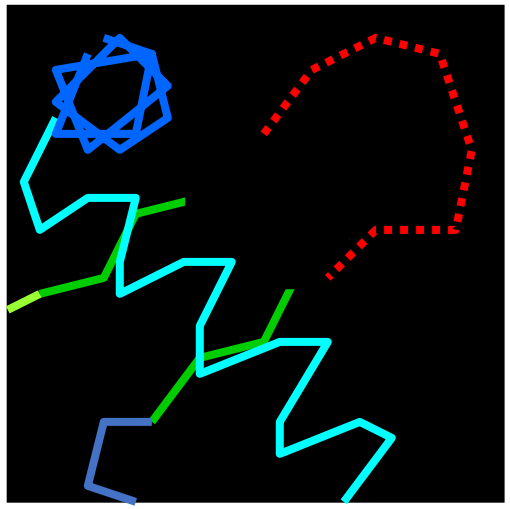
Producing the Main Coordinates

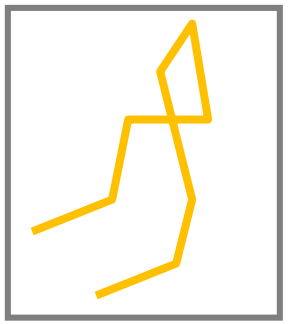
- For matched residues transfer all coordinates to model protein
- For similar residues backbone coordinates are used; locate/replace side chains with proper rotamer
- For different residues only backbone residues are transferred.

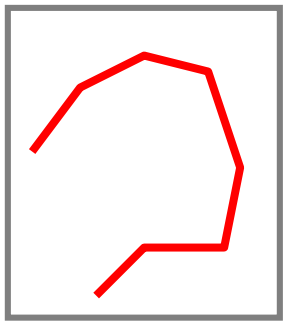
```
Subject: 4dja
ATOM     39  N   ILE A   7       6.036  18.150  68.780
ATOM     40  CA  ILE A   7       6.844  18.371  67.576
ATOM     41  C   ILE A   7       7.921  17.298  67.523
ATOM     42  O   ILE A   7       8.665  17.109  68.491
ATOM     43  CB  ILE A   7       7.471  19.790  67.537
ATOM     45  CG2 ILE A   7       8.373  19.961  66.308
ATOM     46  CD1 ILE A   7       6.771  22.307  67.584
ATOM     47  N   LEU A   8       7.998  16.594  66.395
ATOM     48  CA  LEU A   8       8.991  15.543  66.215
ATOM     49  C   LEU A   8      10.319  16.102  65.721
ATOM     50  O   LEU A   8      10.405  17.260  65.260

Model:
ATOM     71  N   ILE     9       6.524  22.873  68.755
ATOM     72  CA  ILE     9       7.356  22.980  67.596
ATOM     73  CB  ILE     9       8.124  24.269  67.533
ATOM     74  CG1 ILE     9       7.152  25.459  67.486
ATOM     75  CG2 ILE     9       9.063  24.211  66.318
ATOM     76  CD1 ILE     9       7.832  26.810  67.698
ATOM     77  C   ILE     9       8.336  21.849  67.665
ATOM     78  O   ILE     9       8.991  21.636  68.685
ATOM     79  N   LEU    10       8.435  21.077  66.564
ATOM     80  CA  LEU    10       9.312  19.947  66.494
ATOM     81  CB  LEU    10       8.880  18.873  65.493
```

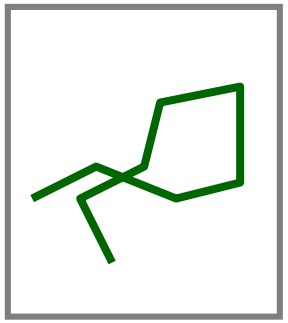
Producing the Coordinates of Variable Regions

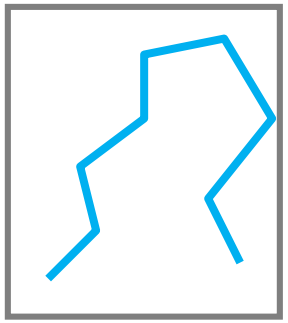
- Finding similar peptide sequences from different proteins
  - All found sequences are safe to use
  - May not fit 100% to given protein model
- De novo structure modelling
- Multiple loop models are offered.
  - Avoid steric overlaps
  - Check atoms within the loop & rest of the protein and the loop
  - Cα-Cα distances should be monitored
  - Locating loops can be performed by using superposition algorithm

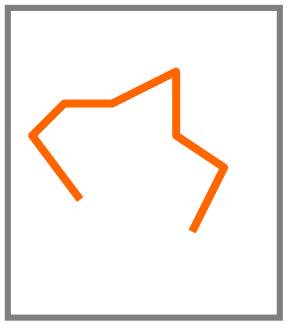


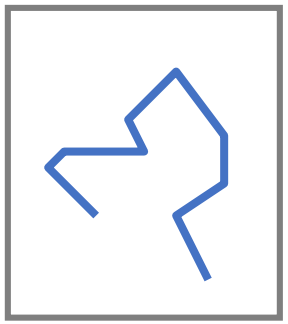




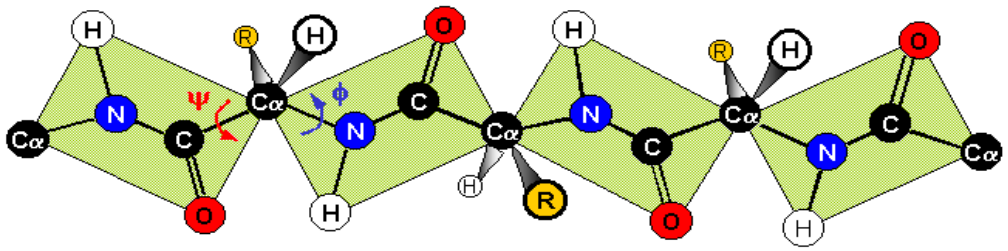


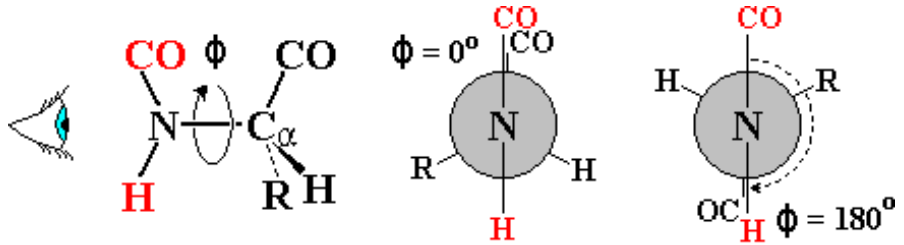


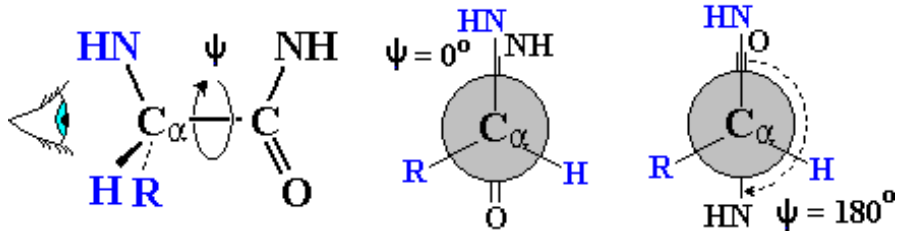


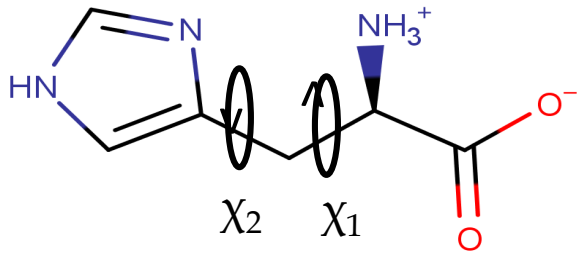


Side Chain Conformation Search

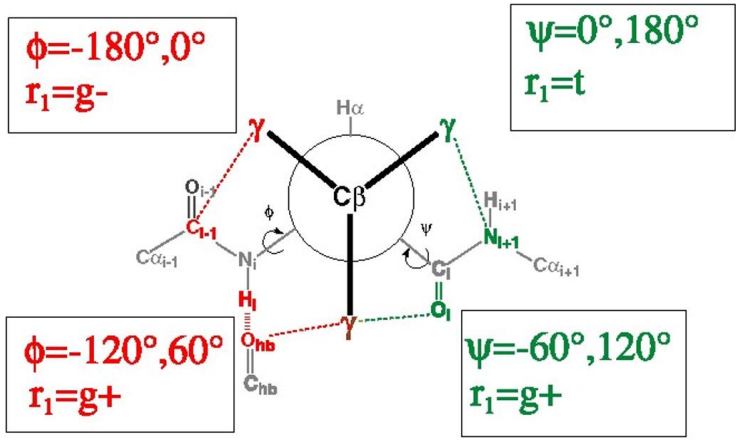


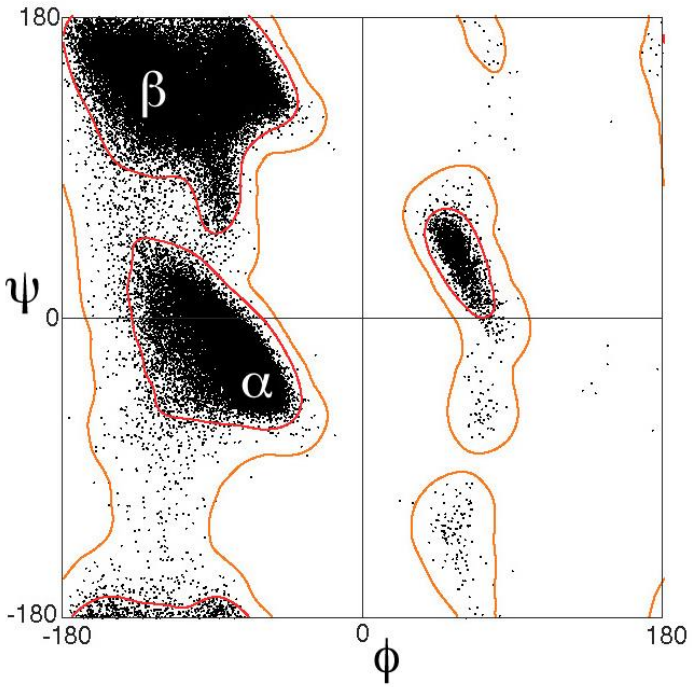




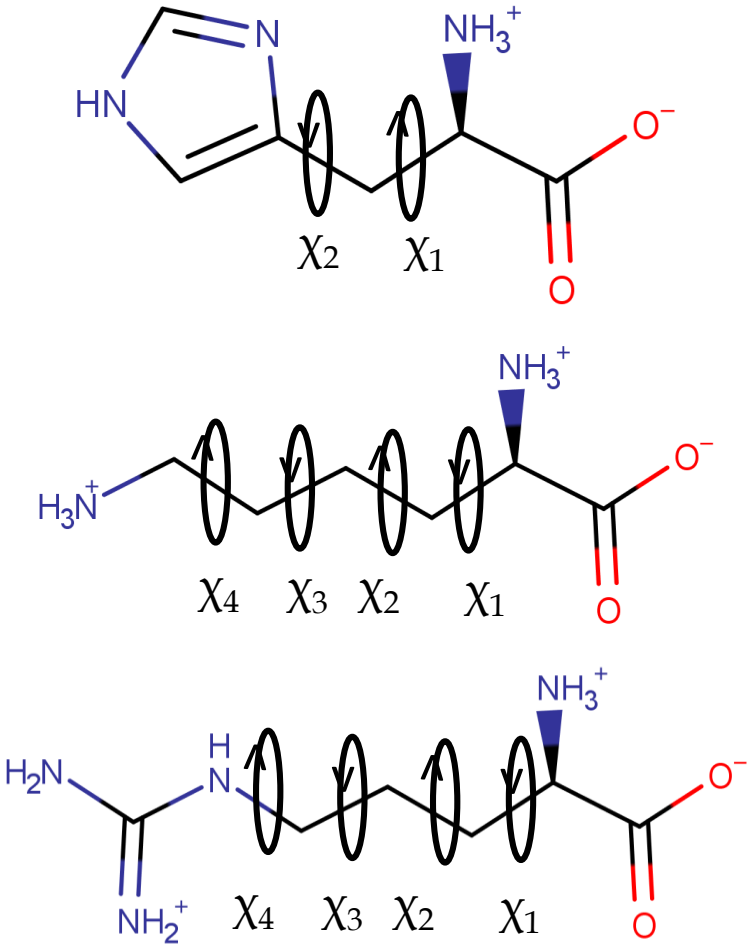


- Some values are favored some are not allowed





Side Chain Conformation Search



- It is hard to find the best conformation of a side chain because of constraints e.g. bond lengths, angles, rotatable backbones
- Some residues have multiple degree of freedom
- Must search for side chain conformations in loops
- Statistical approaches can be utilized to determine correct conformation of side chains
  - Local environment
  - Side chains prefer to adopt conformation to be closely packed

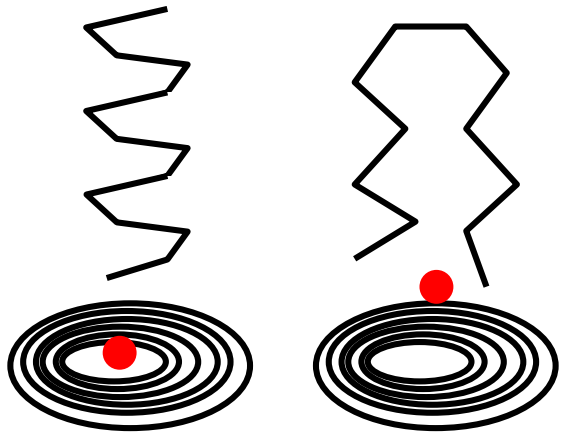
Available Web Servers for Protein Structure Prediction

- SWISS https://swissmodel.expasy.org/
- ModWeb https://modbase.compbio.ucsf.edu/modweb/ (needs registration)
- Phyre2 http://www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index
- RaptorX http://raptorx.uchicago.edu/
- I-TASSER https://zhanglab.ccmb.med.umich.edu/I-TASSER/
- OpenPredict https://open.predictprotein.org/

Refinement of Structure

- No protein model is perfect.
- Substitution of small residues with large ones or vice versa
- Peptide bond can be strained as a result of using multiple templates e.g. for loop modelling
- Conformation of loops are not optimum

- Energy minimization methods can be used to produce conformationally optimum structure
  - e.g. Steepest descent and conjugate gradient methods



- Molecular dynamics is useful to explore various conformational space of proteins

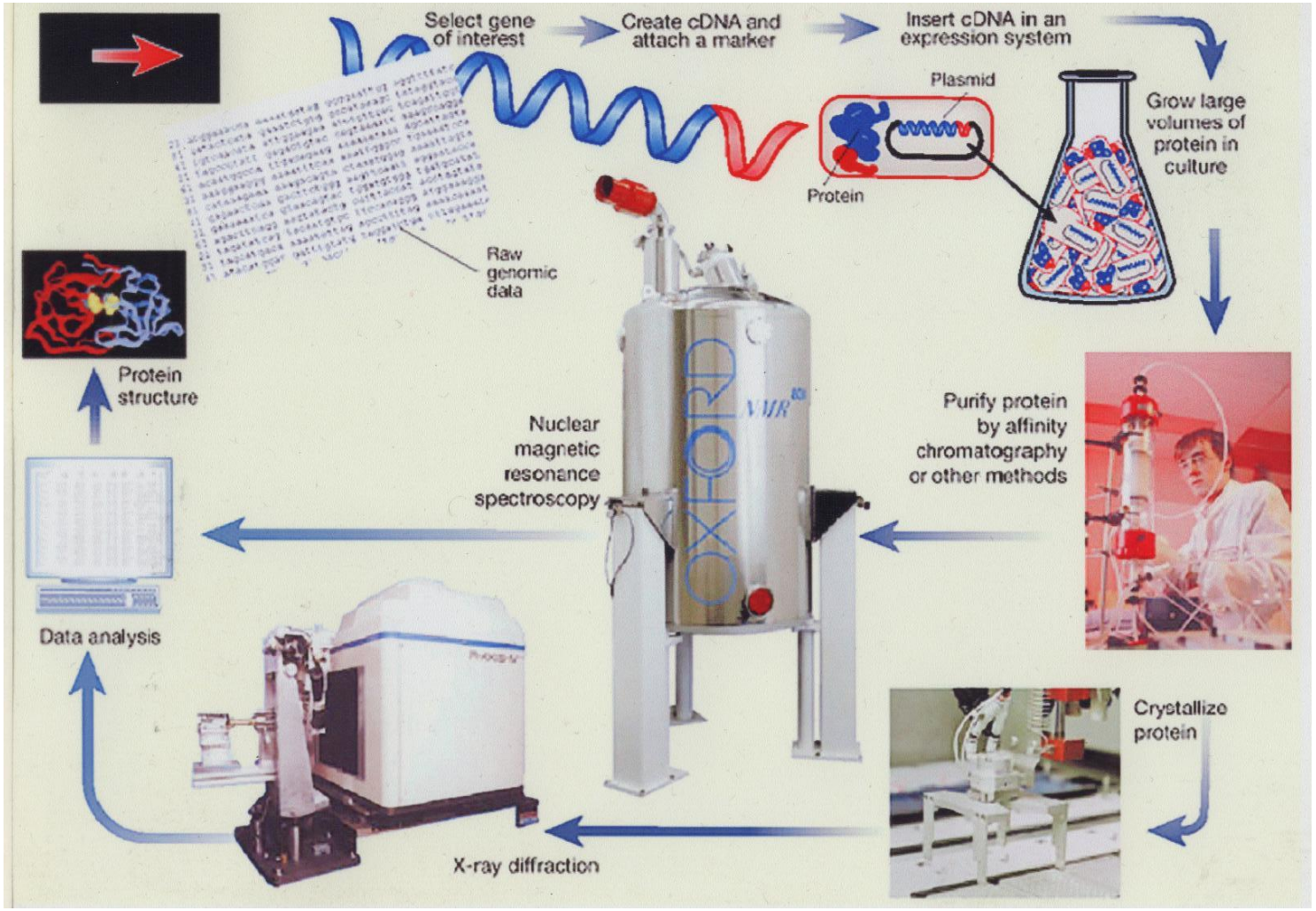
Model Validation

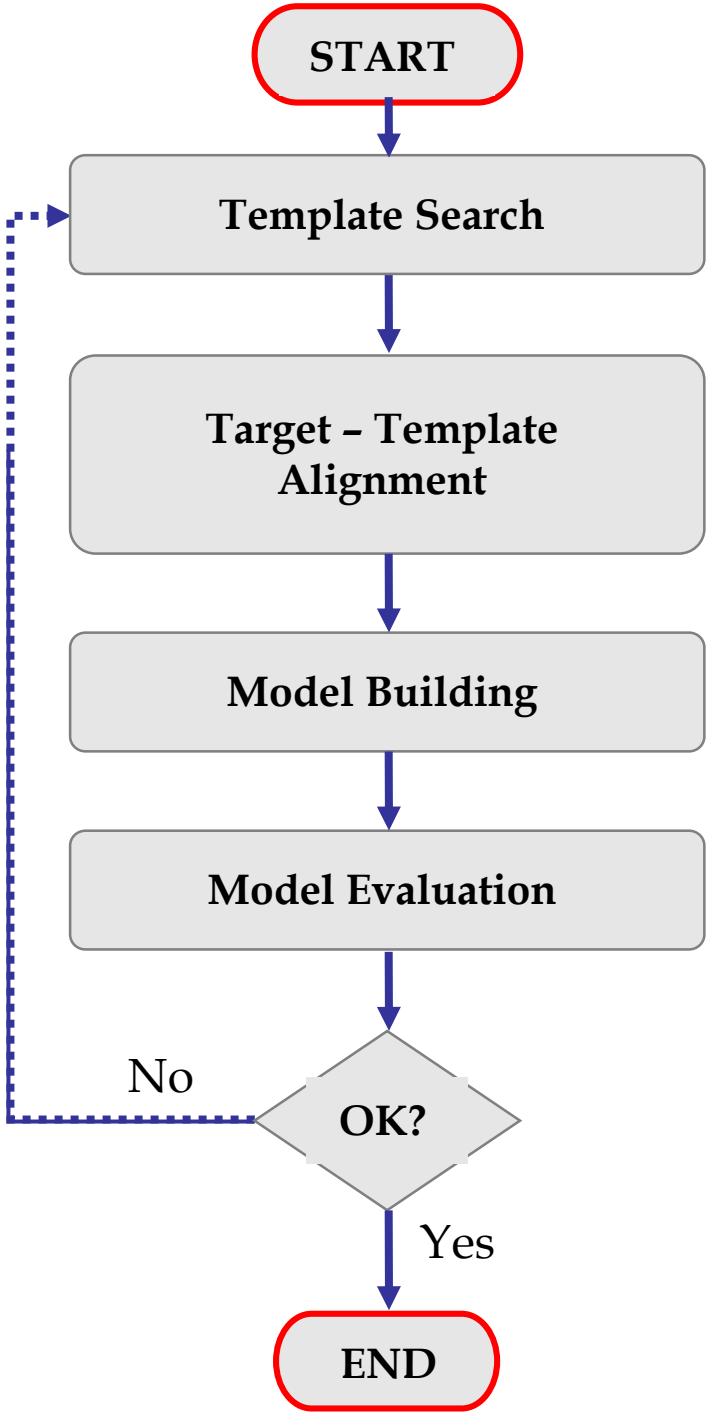
- Quality/correctness of protein model generated via homology modelling is restricted with
  - Sequence similarity
  - Template structure quality
- No model is error free
- Models need to be verified
  - Overall fold e.g. disposition of hydrophobic region
  - Stereochemical properties: bond lengths, angles

- Good protein structures are expected to have
  - Minimum disallowed torsions; interstitial cavities; number of buried charges; radius of gyration; covalent and noncovalent energies;
  - Maximum # of H-bonds; buried hydrophobic residues; exposed hydrophilic residues

Available Web Resources for Structure Validation

- ❖SWISS https://swissmodel.expasy.org/assess/
- ❖ProSA https://prosa.services.came.sbg.ac.at/prosa.php
- ❖ModEval https://modbase.compbio.ucsf.edu/evaluation/
- ❖VERIFY3D https://servicesn.mbi.ucla.edu/Verify3D/
- ❖ERRAT https://servicesn.mbi.ucla.edu/ERRAT/
- ❖WHAT IF https://swift.cmbi.umcn.nl/servers/html/index.html
- ❖WHATCHECK https://servicesn.mbi.ucla.edu/WHATCHECK/
- ❖PROVE https://servicesn.mbi.ucla.edu/PROVE/
- ❖MOLPROBITY http://molprobity.biochem.duke.edu/
- ❖Prediction Tools http://crdd.osdd.net/pstr.php


Select gene of interest
Create cDNA and attach a marker
Insert cDNA in an expression system
Plasmid
Protein
Grow large volumes of protein in culture
Raw genomic data
Protein structure
Nuclear magnetic resonance spectroscopy
Purify protein by affinity chromatography or other methods
Data analysis
Crystallize protein
X-ray diffraction


START
Template Search
Target – Template Alignment
Model Building
Model Evaluation
OK?
No
Yes
END

THANK YOU FOR YOUR ATTENTION