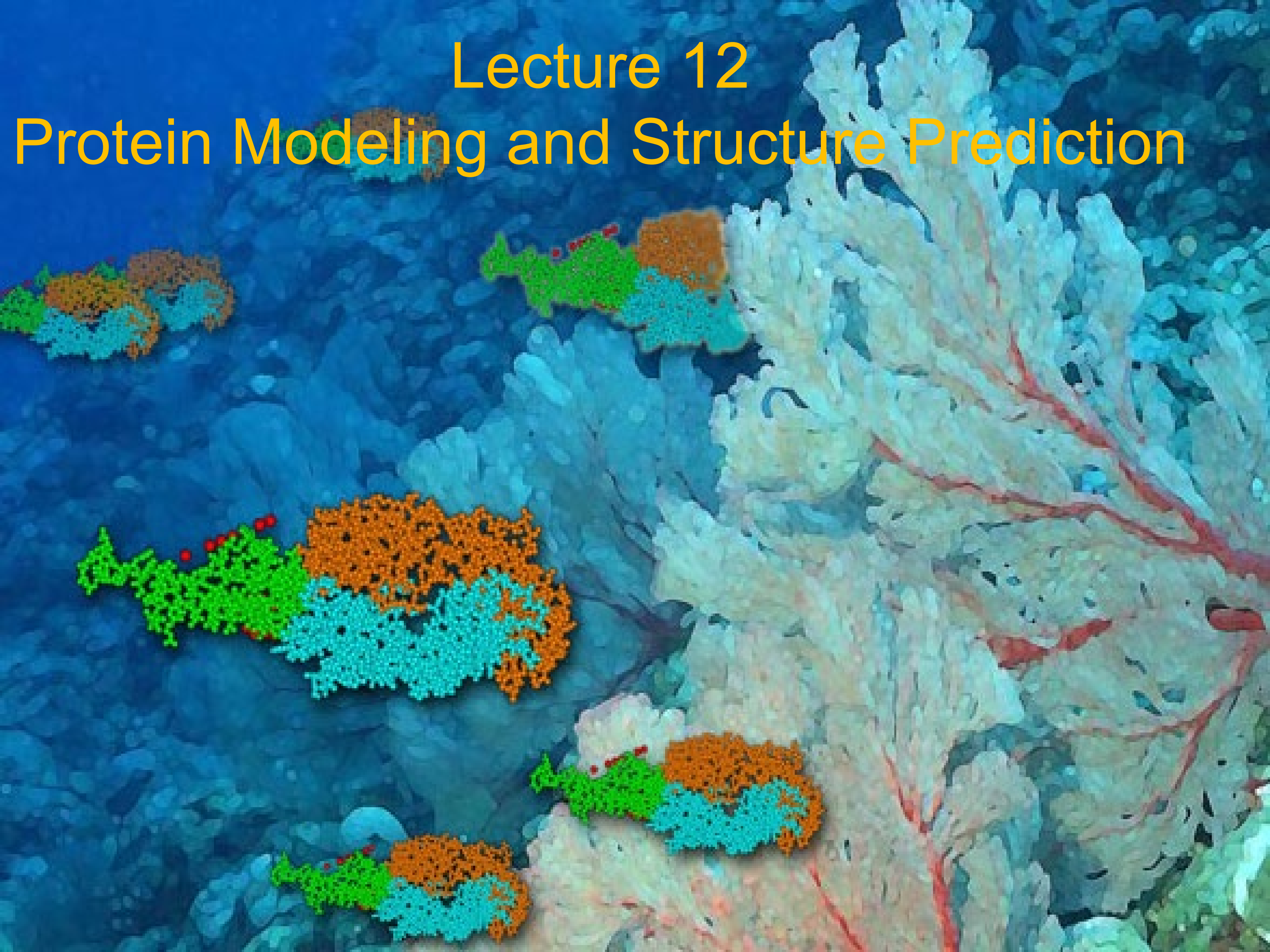
Lecture 12
Protein Modeling and Structure Prediction



Overview

1. Basic molecular modeling
2. Protein structure prediction
   - Secondary
   - 3-D
3. Summarization

1. Molecular modeling
How does a protein fold?

- Most newly synthesized proteins fold without assistance!
  - *Ribonuclease A: denatured protein could refold and recover its activity (C. Anfinsen - 1966)*
- *The amino acid sequence encodes the protein's structural information!*
- *Protein sequence→ structure → function*

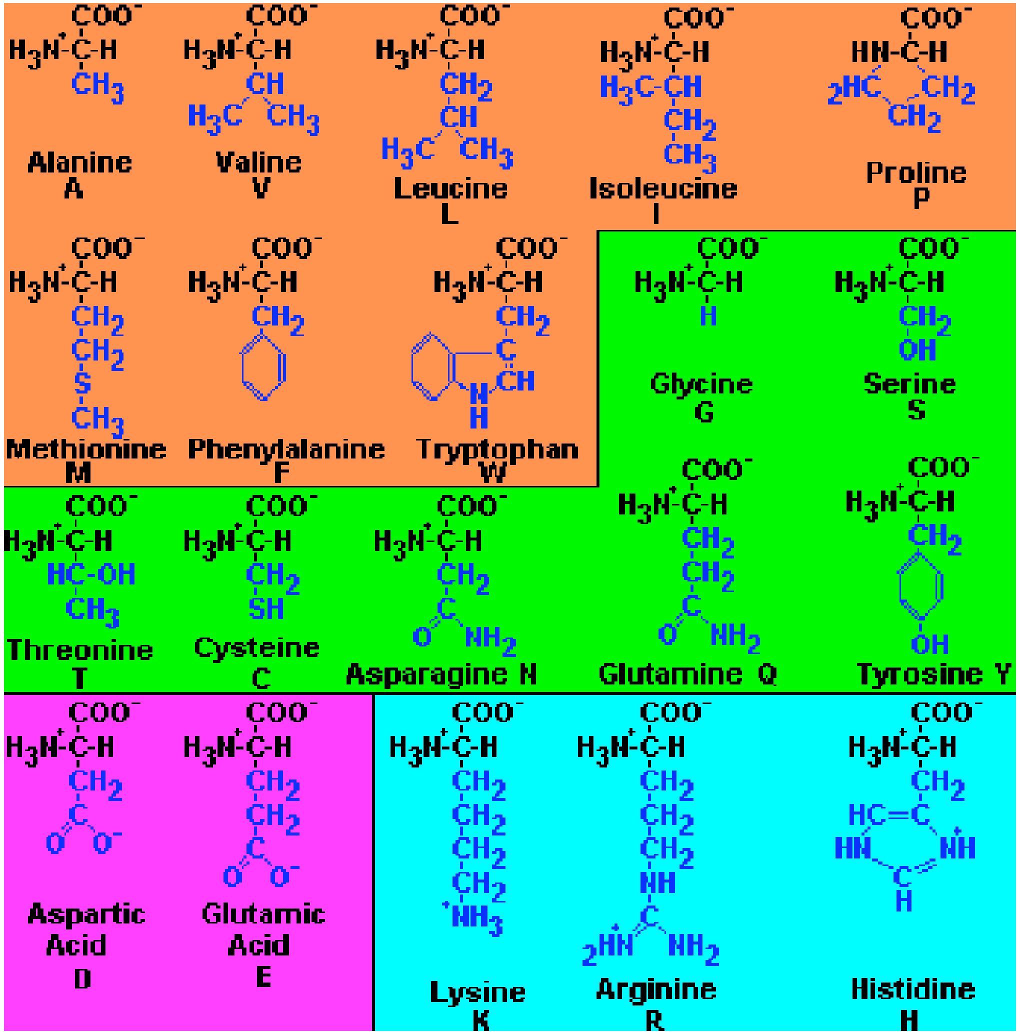
Types of amino acids

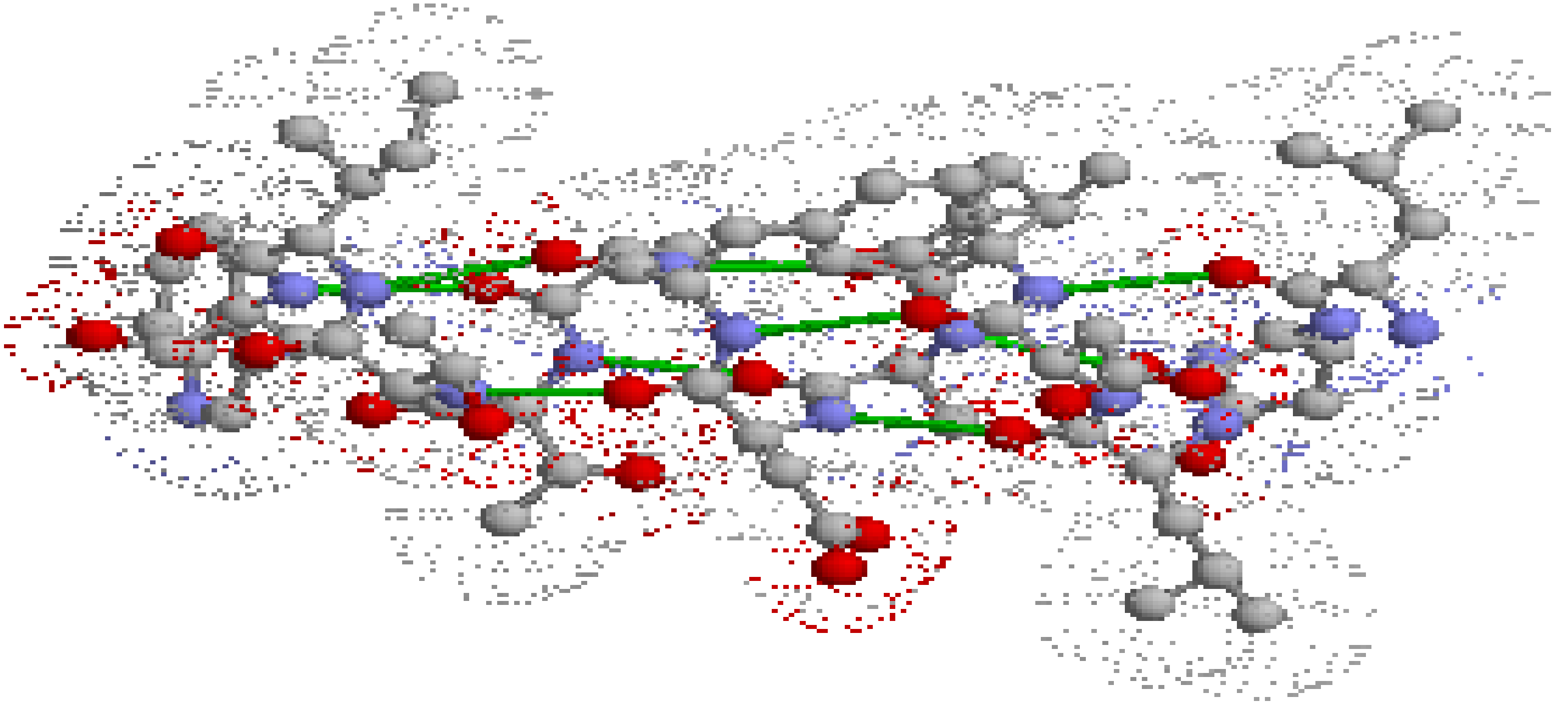
Hydrophobic

Hydrophilic, Neutral

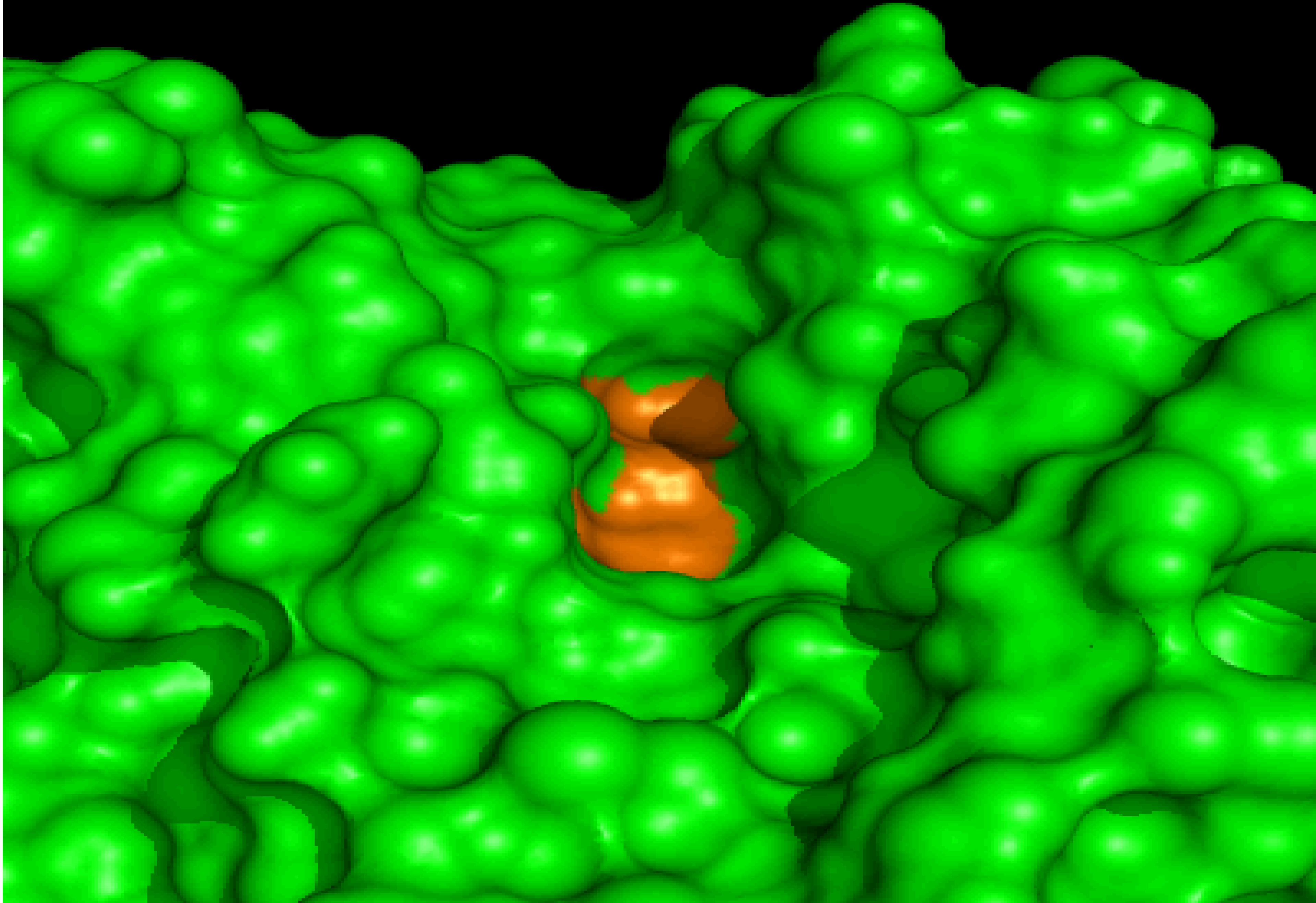
Hydrophilic, Acidic

Hydrophilic, Basic

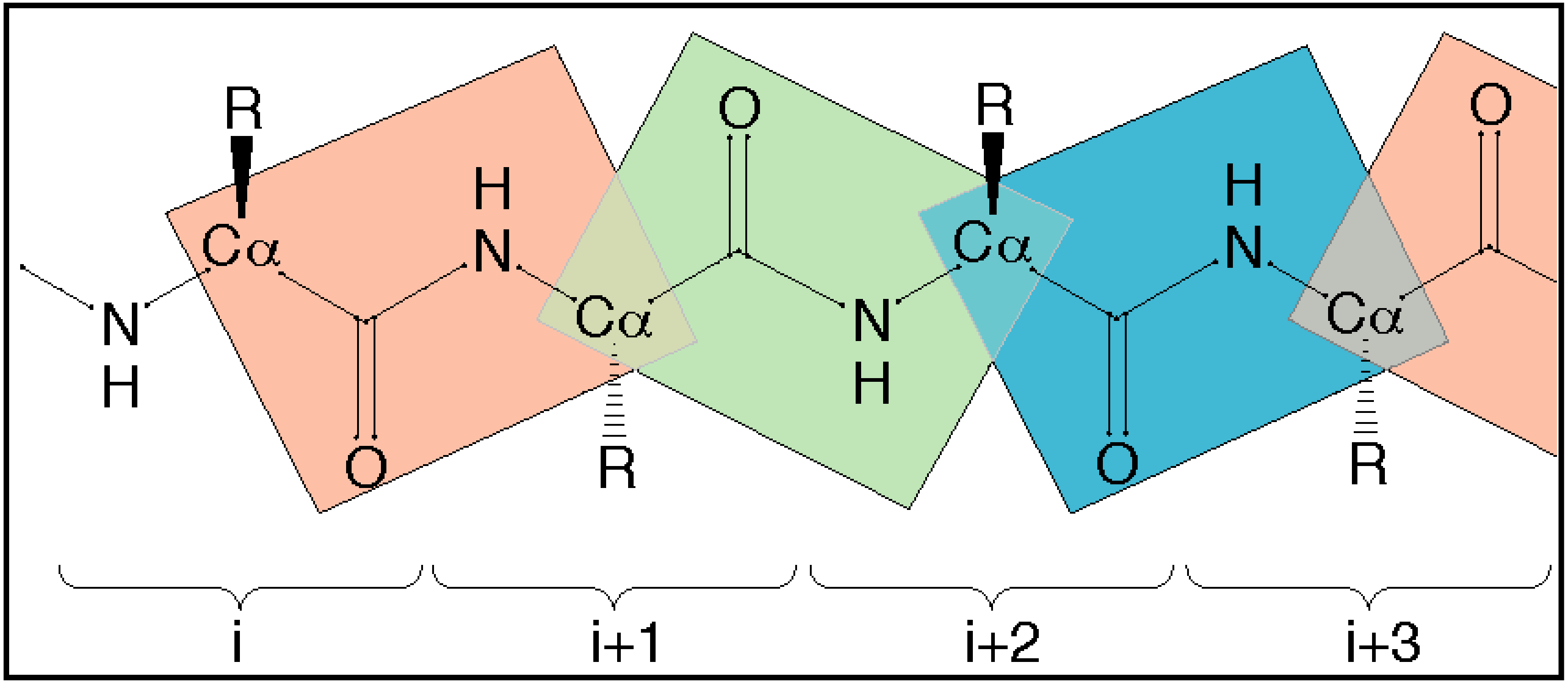




Molecular modeling methods are the theoretical methods and computational techniques used to simulate the behavior of molecules and molecular systems

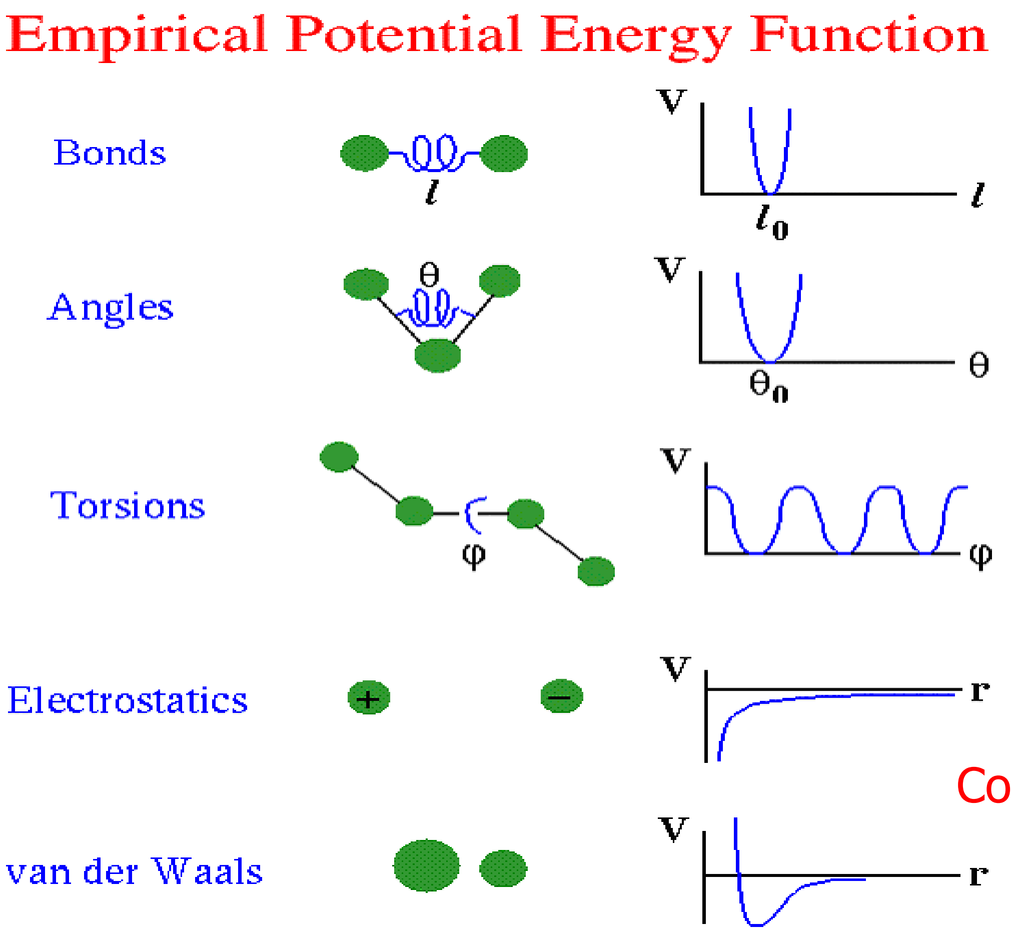


Peptide Chain



Atoms interaction?

Molecular Modeling:
Basic Interactions and Their Models



Covalent: Bond stretching + angle bending + torsion

Non-covalent: Electrostatics+Van der Waals + Hydrogen bond

Potential Energy Function (PEF) and force fields

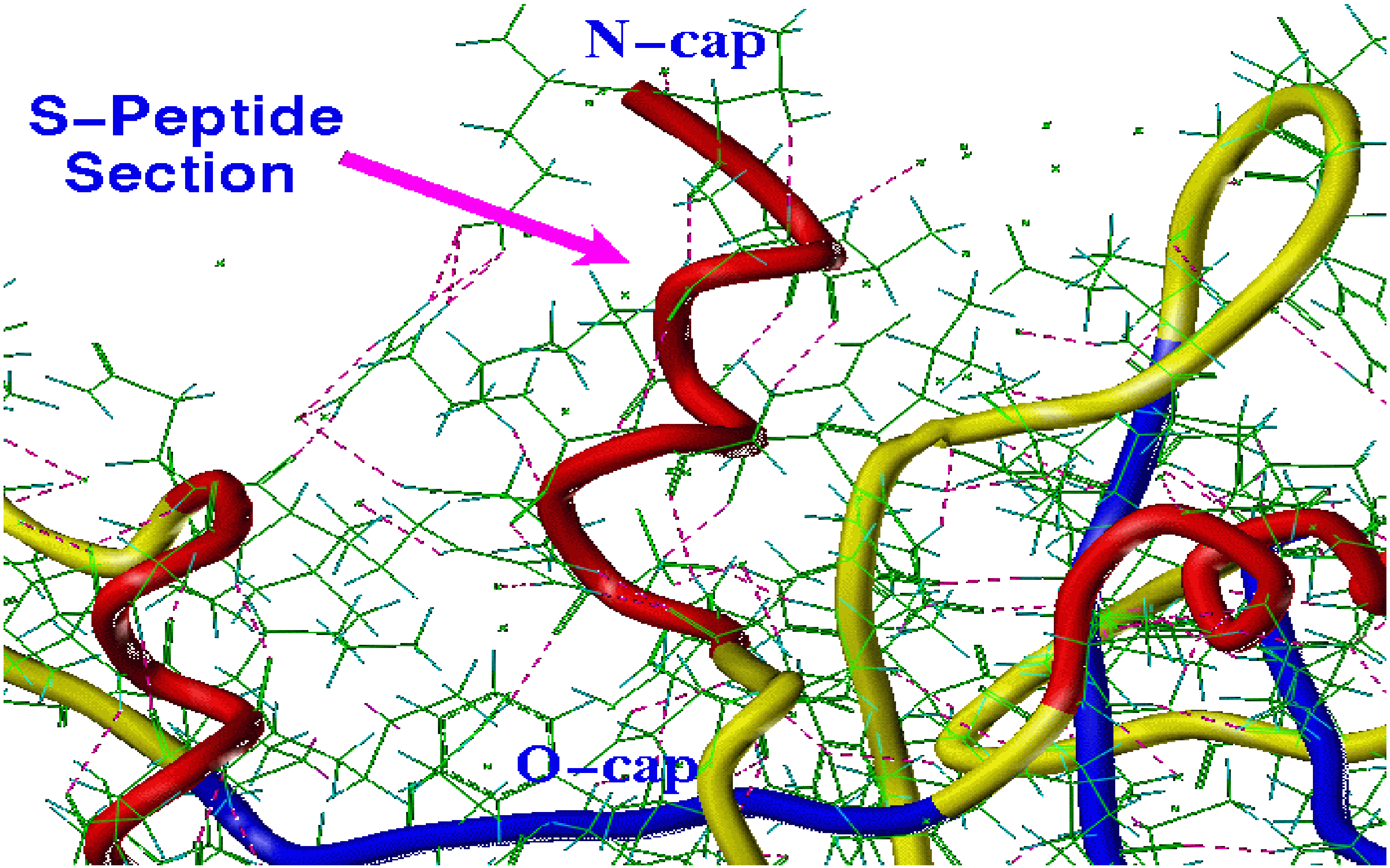
- PEF is the total potential energy which is defined as the energy difference between a real and an ideal molecule

$$PEF(R) = \sum_{bond-stretch} \frac{1}{2} k_r (r - r_{eq})^2 + \sum_{bond-angle-bending} \frac{1}{2} k_\theta (\theta - \theta_{eq})^2 +$$

$$\sum_{bond-rotation} \frac{v_n}{2} [1 + \cos(n\phi - \gamma)] + \sum_{H-bond} [V_0 (1 - e^{-a(r-r_0')})^2 - V_0] + \sum_{non-bonded} [\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} + \frac{q_i q_j}{\varepsilon_{ij} r_{ij}}]$$

- Force fields: Amber, Charmm, Gromos etc.

Molecular Modeling: Hydrogen Bond

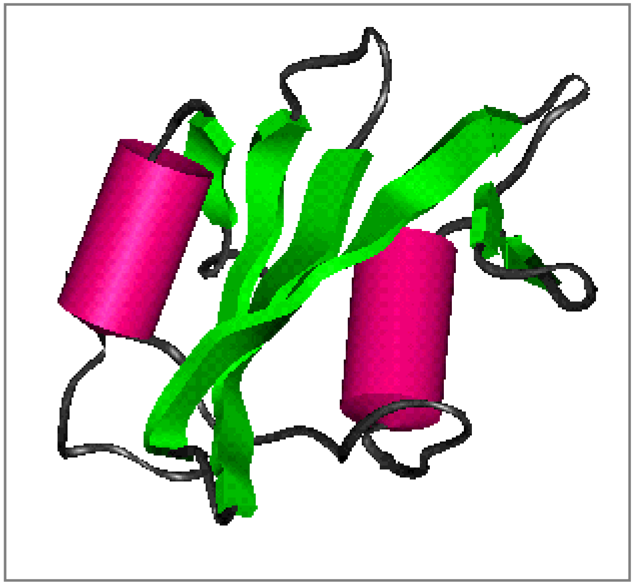


$$
\begin{aligned}
V_H(r) &= A/r^{12} - B/r^{6} + q_i q_j/\varepsilon_r r_{ij} && \text{for AMBER} \\
&= (A/r^{12} - B/r^{10})\cos^m(\theta_{A\text{-}H\text{-}D})\cos^n(\theta_{AA\text{-}A\text{-}H})sw_1(\mathbf{r})sw_2(\theta) && \text{for CHARM} \\
&= V_0\,(1-e^{-a(r-r0)})^2 - V_0 && \text{for Prohofsky/Chen}
\end{aligned}
$$

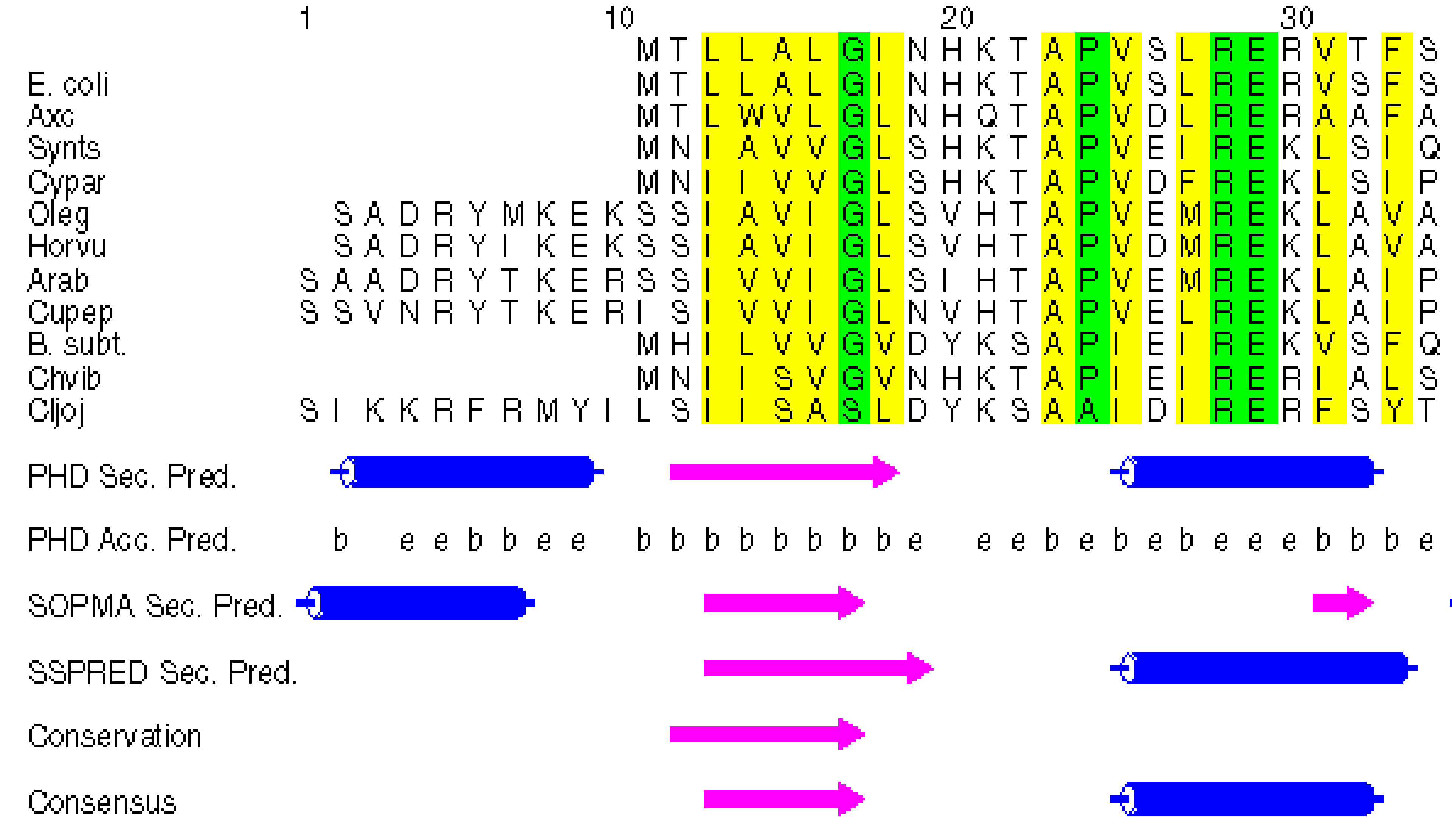
2. Structure prediction

MHGAYRTPRSKTDAYGCQILETRAS

???



Secondary structure prediction



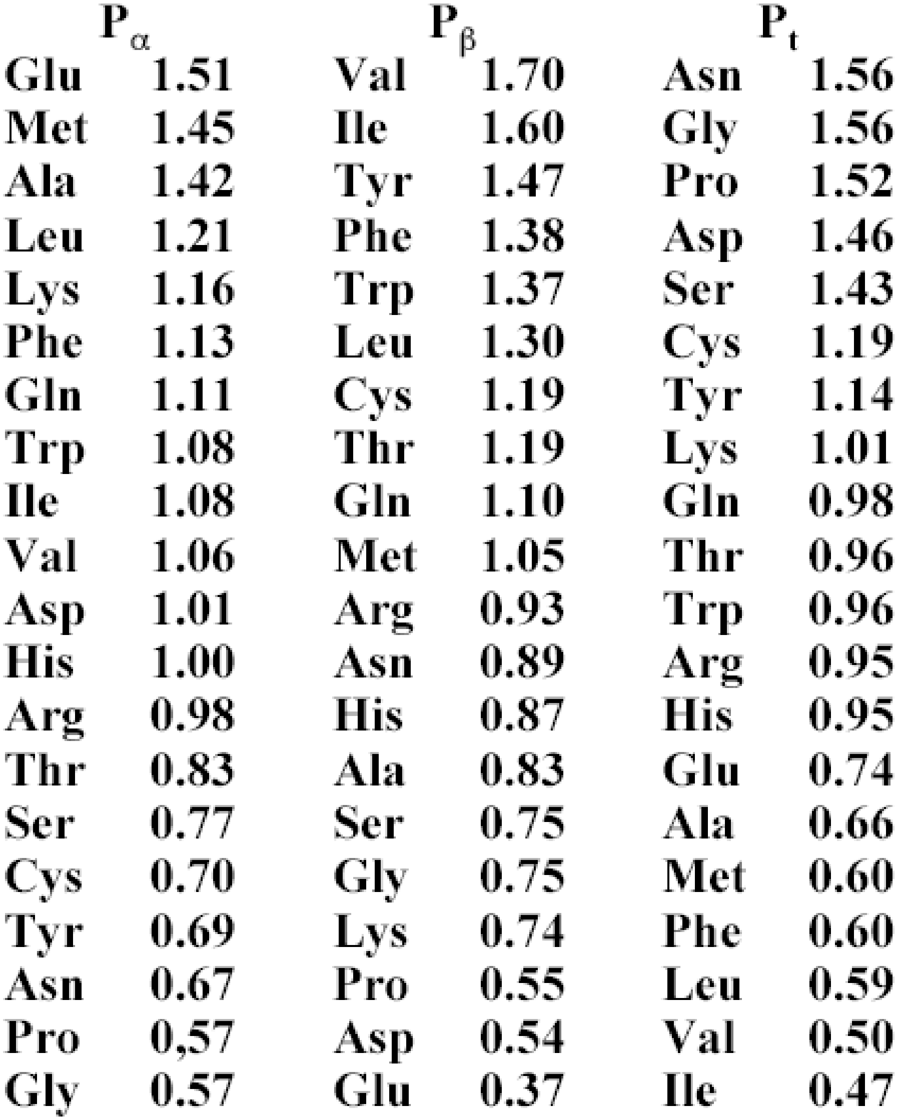
```
                  1                 10                  20                  30
                                      M T L L A L G I N H K T A P V S L R E R V S F S
E. coli                               M T L L A L G I N H K T A P V S L R E R V S F S
Axc                                   M T L W V L G L N H Q T A P V D L R E R A A F A
Synts                                 M N I A V V G L S H K T A P V E I R E K L S I Q
Cypar                                 M N I I V V G L S H K T A P V D F R E K L S I P
Oleg                S A D R Y M K E K S S I A V I G L S V H T A P V E M R E K L A V A
Horvu               S A D R Y I K E K S S I A V I G L S V H T A P V D M R E K L A V A
Arab              S A A D R Y T K E R S S I V V I G L S I H T A P V E M R E K L A I P
Cupep             S S V N R Y T K E R I S I V V I G L N V H T A P V E L R E K L A I P
B. subt.                              M H I L V V G V D Y K S A P I E I R E K V S F Q
Chvib                                 M N I I S V G V N H K T A P I E I R E R I A L S
Cljoj             S I K K R F R M Y I L S I I S A S L D Y K S A A I D I R E R F S Y T

PHD Sec. Pred.

PHD Acc. Pred.      b   e e b b e e   b b b b b b b b e   e e b e b e b e e e b b b e

SOPMA Sec. Pred.

SSPRED Sec. Pred.

Conservation

Consensus
```

Protein Secondary Structure Prediction:

Structural propensity of amino acids

Each residue is assigned to one of the three classes:

- Forming residues – favor a structure
- Indifferent residues
- Breaking residues – stop the extension of a structure



P_α		P_β		P_t	
Glu	1.51	Val	1.70	Asn	1.56
Met	1.45	Ile	1.60	Gly	1.56
Ala	1.42	Tyr	1.47	Pro	1.52
Leu	1.21	Phe	1.38	Asp	1.46
Lys	1.16	Trp	1.37	Ser	1.43
Phe	1.13	Leu	1.30	Cys	1.19
Gln	1.11	Cys	1.19	Tyr	1.14
Trp	1.08	Thr	1.19	Lys	1.01
Ile	1.08	Gln	1.10	Gln	0.98
Val	1.06	Met	1.05	Thr	0.96
Asp	1.01	Arg	0.93	Trp	0.96
His	1.00	Asn	0.89	Arg	0.95
Arg	0.98	His	0.87	His	0.95
Thr	0.83	Ala	0.83	Glu	0.74
Ser	0.77	Ser	0.75	Ala	0.66
Cys	0.70	Gly	0.75	Met	0.60
Tyr	0.69	Lys	0.74	Phe	0.60
Asn	0.67	Pro	0.55	Leu	0.59
Pro	0,57	Asp	0.54	Val	0.50
Gly	0.57	Glu	0.37	Ile	0.47

Protein Secondary Structure Prediction:

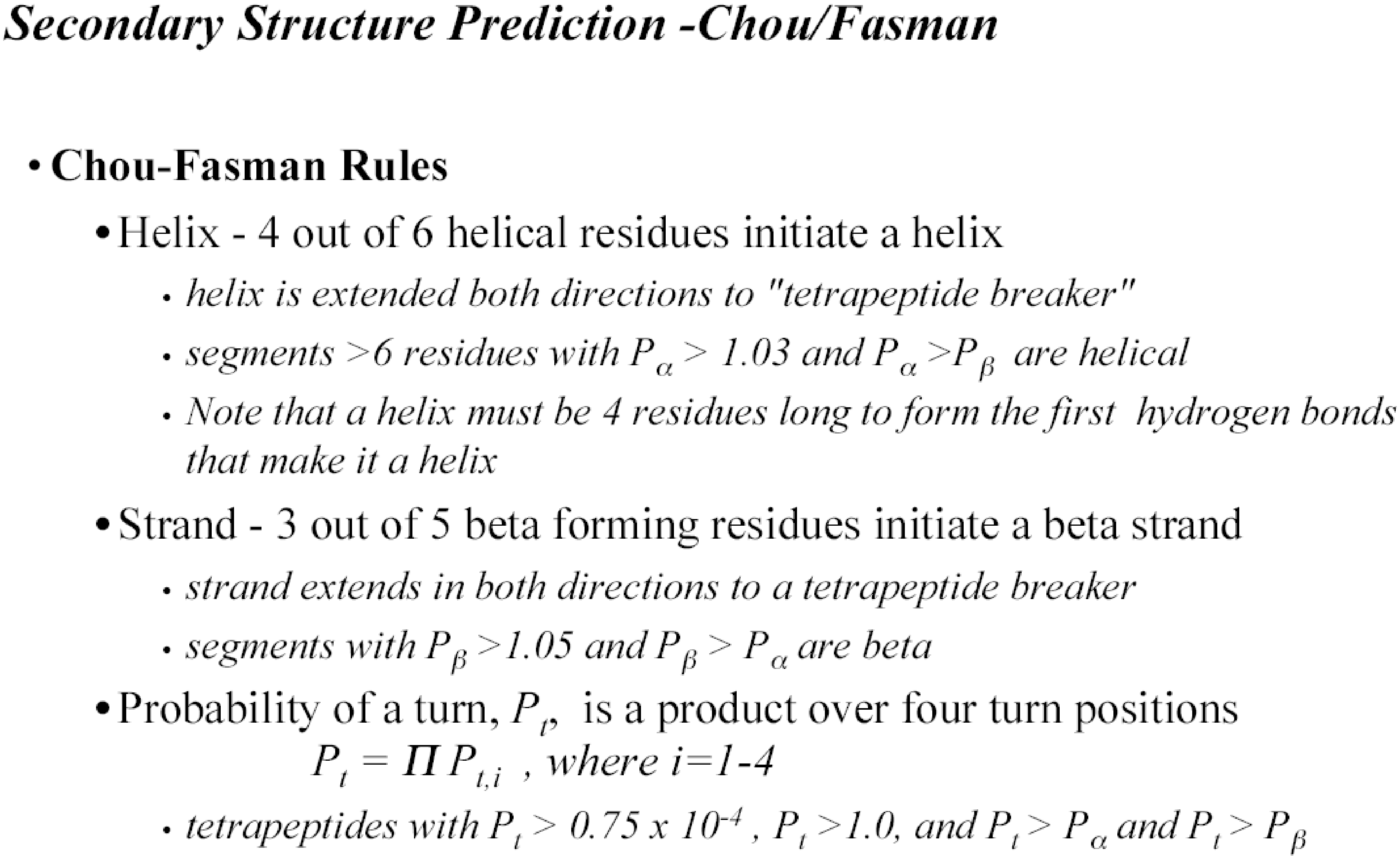
Chou and Fasman procedure

- Find helical initiation regions
- Extend helices until they reach tetrapeptide breakers
- Find beta initiation regions
- Extend until they reach tetrapeptide breakers
- Find turns
- Resolve conflicts between alpha and beta

Chou and Fasman did not provide an explicit algorithm for this conflict resolution, relying on their expert judgment. This meant that each person's prediction could be different. Most people are not experts.

"Prediction of the secondary structure of proteins from their amino acid sequence",
P. Y. Chou, G. D. Fasman, 1978, *Adv. Enzymolog. Relat. Areas Mol. Biol.*, 47, 45-147.

Protein Secondary Structure Prediction:



Secondary Structure Prediction -Chou/Fasman

- **Chou-Fasman Rules**
  - Helix - 4 out of 6 helical residues initiate a helix
    - *helix is extended both directions to "tetrapeptide breaker"*
    - *segments >6 residues with $P_\alpha > 1.03$ and $P_\alpha > P_\beta$ are helical*
    - *Note that a helix must be 4 residues long to form the first hydrogen bonds that make it a helix*
  - Strand - 3 out of 5 beta forming residues initiate a beta strand
    - *strand extends in both directions to a tetrapeptide breaker*
    - *segments with $P_\beta > 1.05$ and $P_\beta > P_\alpha$ are beta*
  - Probability of a turn, P_t, is a product over four turn positions

    $$P_t = \Pi P_{t,i} \text{ , where } i=1\text{-}4$$

    - *tetrapeptides with $P_t > 0.75 \times 10^{-4}$, $P_t > 1.0$, and $P_t > P_\alpha$ and $P_t > P_\beta$*

Some softwares for secondary structure prediction :

- PHD or PredictProtein: Rost and Sander - (http://www.embl-heidelberg.de/predictprotein/predictprotein.html)
- *JPRED: Cuff and Barton - (http://circinus.ebi.ac.uk:8081)*
- PREDATOR: Frishman and Argos - (http://www.embl-heidelberg.de/argos/)

pause

Protein 3-D structure prediction

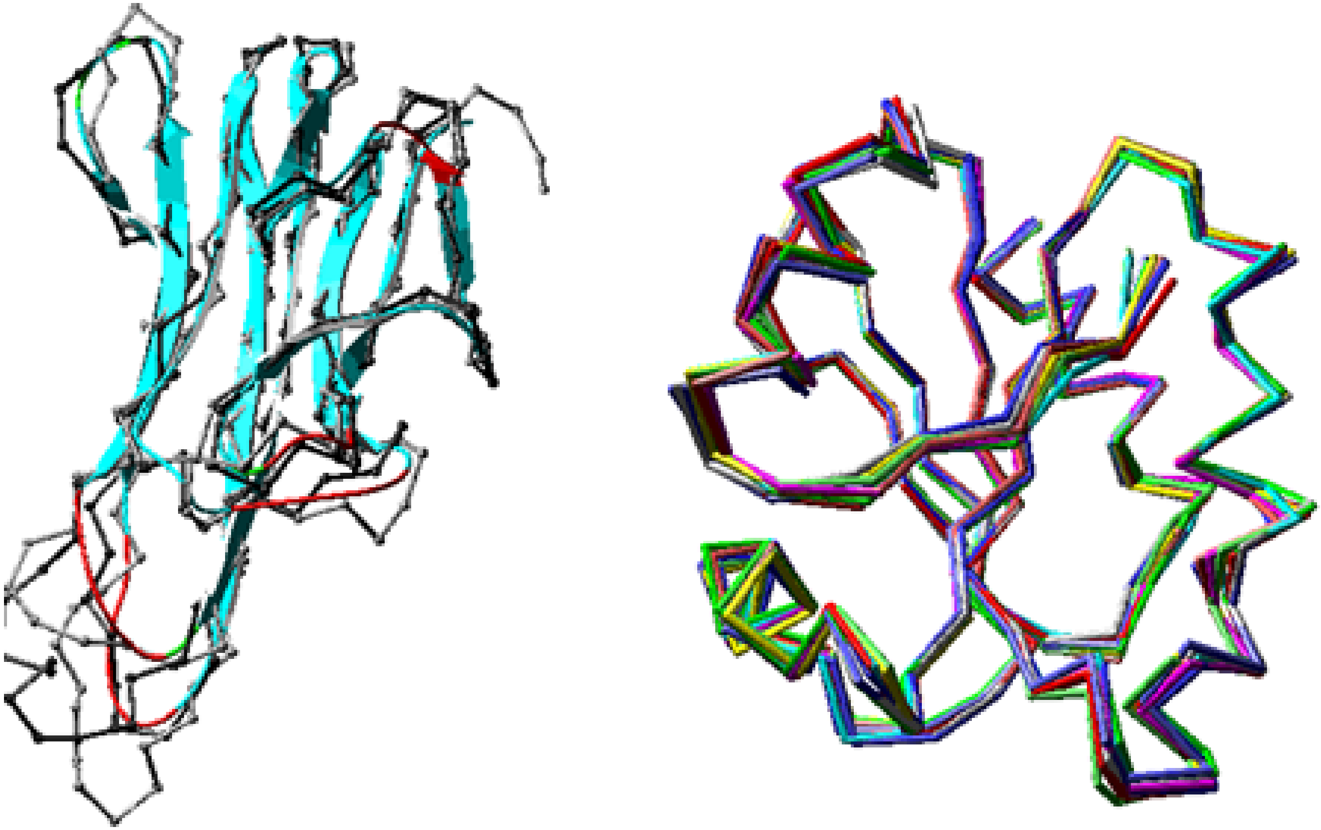
1. Homologue modeling
   - Swiss-Model - an automated homology modeling server developed at Glaxo-Welcome Experimental Research in Geneva.
   - http://www.expasy.ch/swissmod/

2. Threading
   - The problem of aligning a protein sequence to a given structural model is known as protein threading.

3. Ab initio Methods:
   - **ab initio** means from the beginning.
   - MD and Simplified models

Homology Modeling:

Homology models can be very smart!

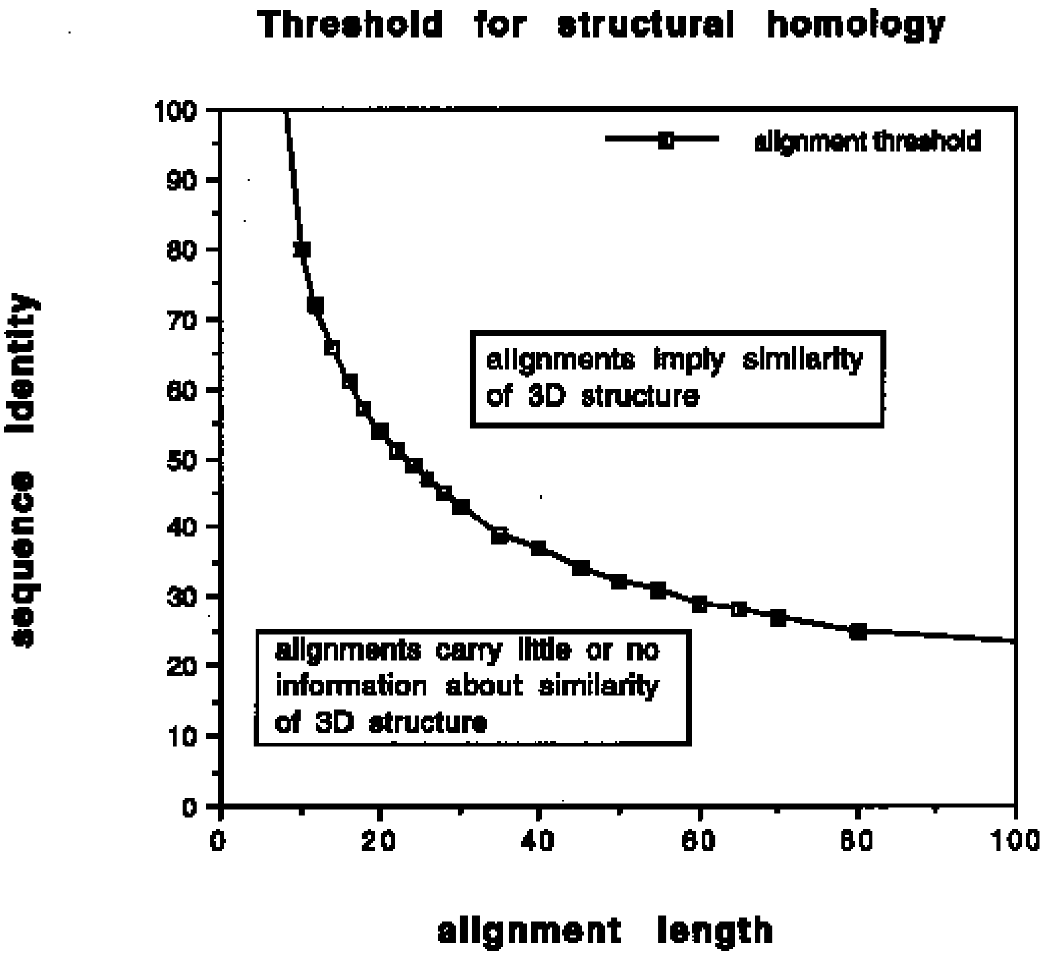


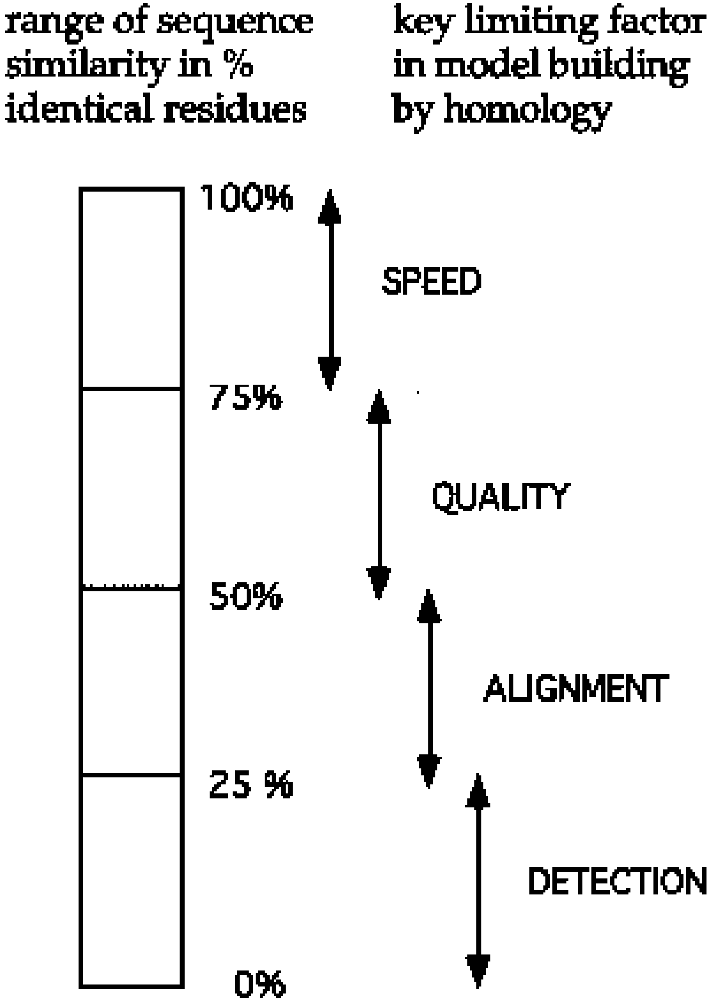
Homology Modeling:

Basic Idea:

- Similar sequence=> Similar structure
- Structure is conserved more than sequence
- Structure of new protein derived using existing protein structures as templates.
- Changes are compensated for locally.

Homology Modeling:





range of sequence similarity in % identical residues

key limiting factor in model building by homology

100%

75%

50%

25 %

0%

SPEED

QUALITY

ALIGNMENT

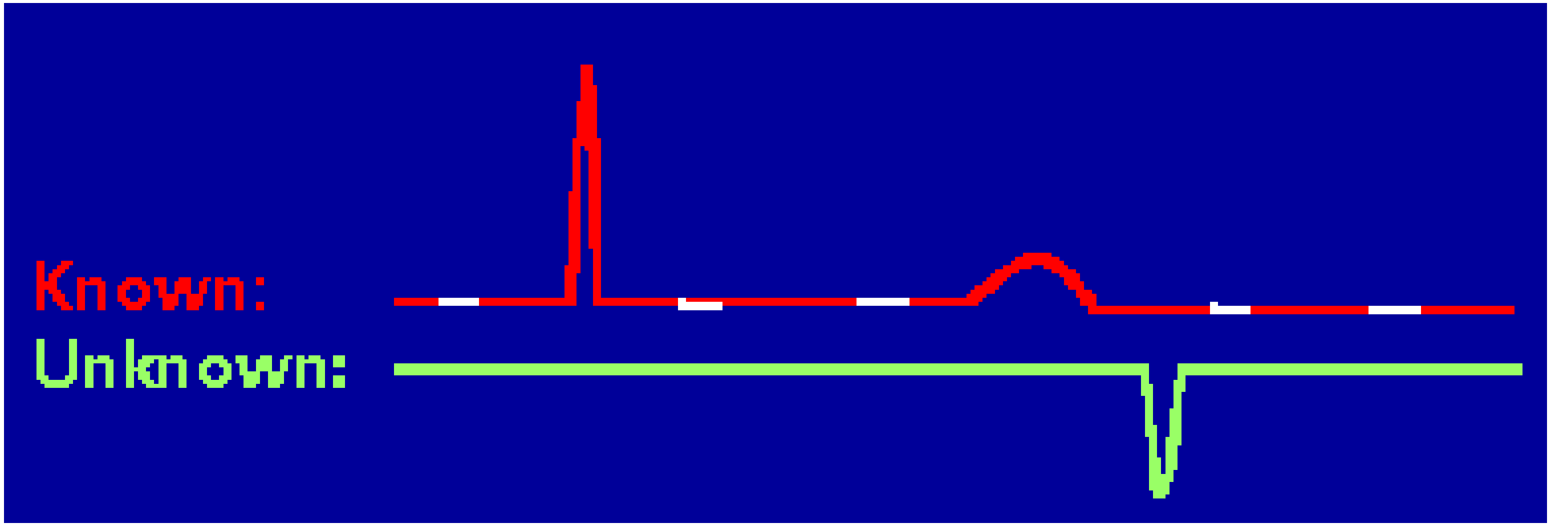
DETECTION

Twilight Zone: below 25% sequence homology

Homology Modeling:

Step One:

- Align sequence of your protein (unknown) with that of candidate template proteins (known)



Searching for homologous template(s)

- Use sequence alignment search programs (e.g., BLAST) to identify homologous sequences with known 3D structure

- Profile-based search methods (e.g., PSI-BLAST) are commonly used to detect weak homologous sequences

- Fold recognition methods can be used to detect potential templates with very weak or nonexistent sequence homology (% identity < 25%)

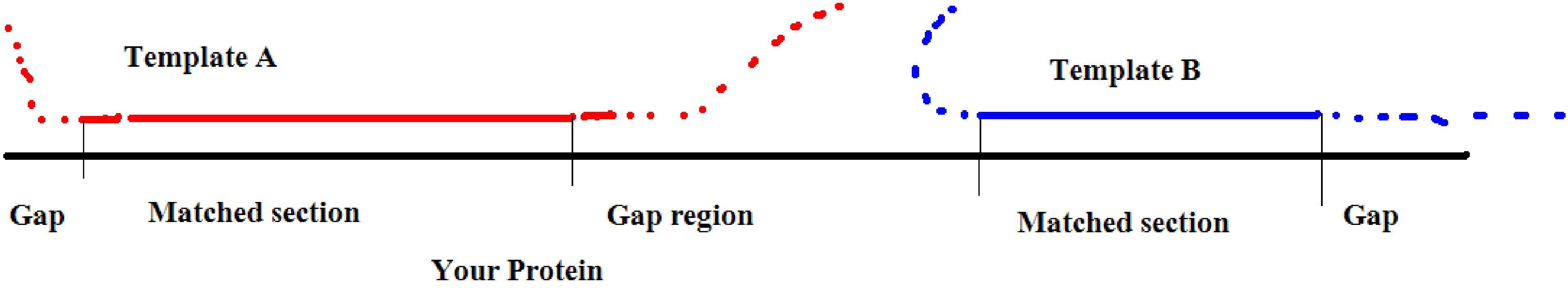
- One approach is to select the template with the highest sequence identity or best alignment to the protein sequence

- Alternatively, multiple templates can be used to construct the model
  - Select a potentially different template for each segment of the protein sequence (based on sequence similarity, etc)
- Other factors in template selection:
  - Resolution of template structure: better to use high resolution structures as model templates
  - Other sources of similarity between protein sequence and template (e.g. similar function, ligands, environment, etc.)

Homology Modeling:

Step Two: Aligning protein sequence with templates

- Select template proteins based on sequence similarity and minimize their X-ray structures
- The whole sequence can be matched by one or more templates



Alignment:

- Can use common pair-wise and multiple sequence Alignment tools to align sequence to template(s)
- Alignment of the sequence to the template can be particularly challenging for homologous templates with low sequence identity (<30-40% identity)
- The accuracy of the alignment of the sequence to template(s) is often the critical parameter for successful homology modeling
  - If sequences are aligned incorrectly, the model will be inaccurate or wrong

Alignment: Challenges

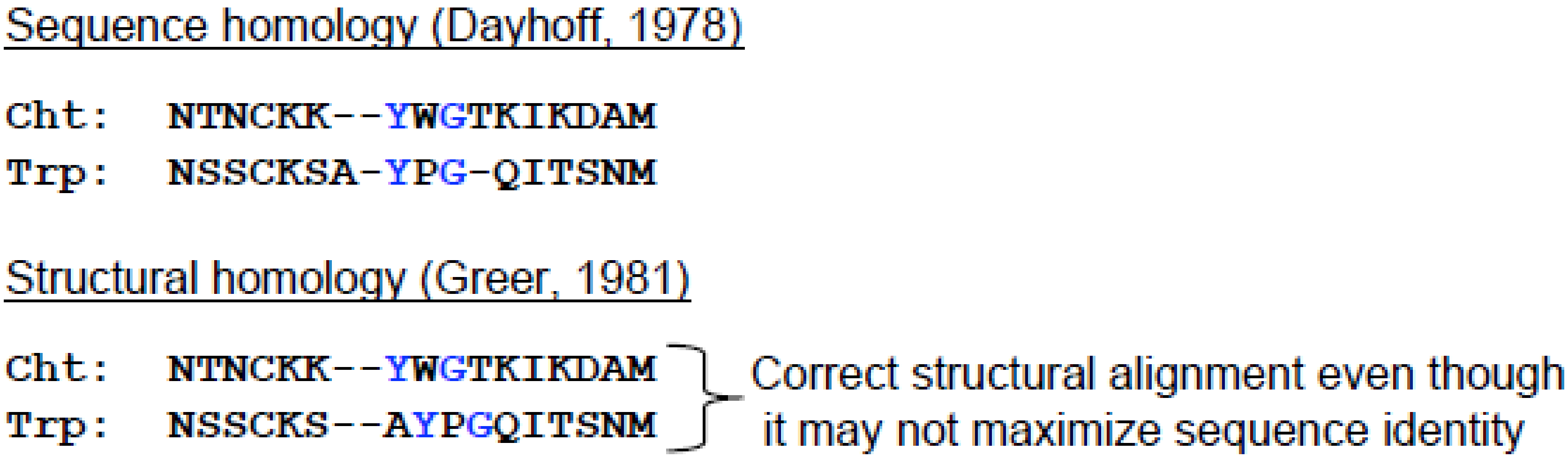
- Typical alignment methods try to maximize sequence similarity (or the score) of the alignment

- However, in homology modeling, we are trying to maximize the structural similarity of residues in our alignment

- Sequence and structural homology are often related, but not always

Example:

chymotrypsin (Cht) and trypsin (Trp)



Sequence homology (Dayhoff, 1978)

```
Cht:   NTNCKK--YWGTKIKDAM
Trp:   NSSCKSA-YPG-QITSNM
```

Structural homology (Greer, 1981)

```
Cht:   NTNCKK--YWGTKIKDAM
Trp:   NSSCKS--AYPGQITSNM
```

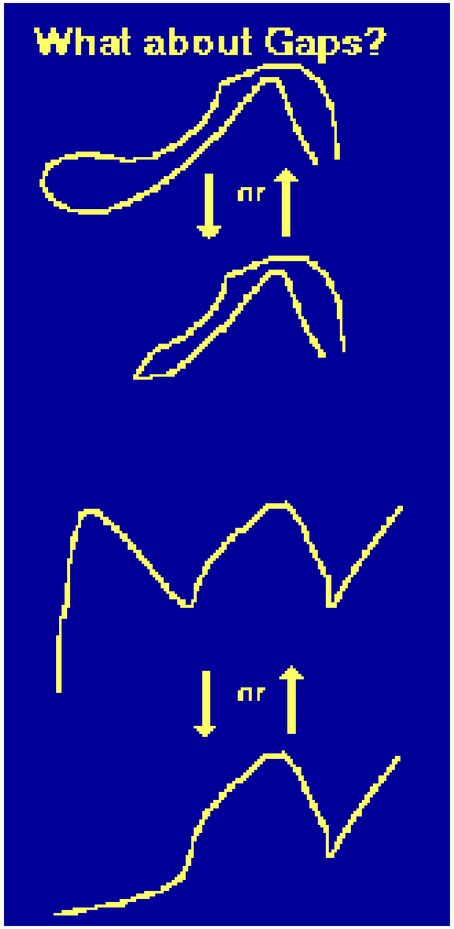
Correct structural alignment even though it may not maximize sequence identity

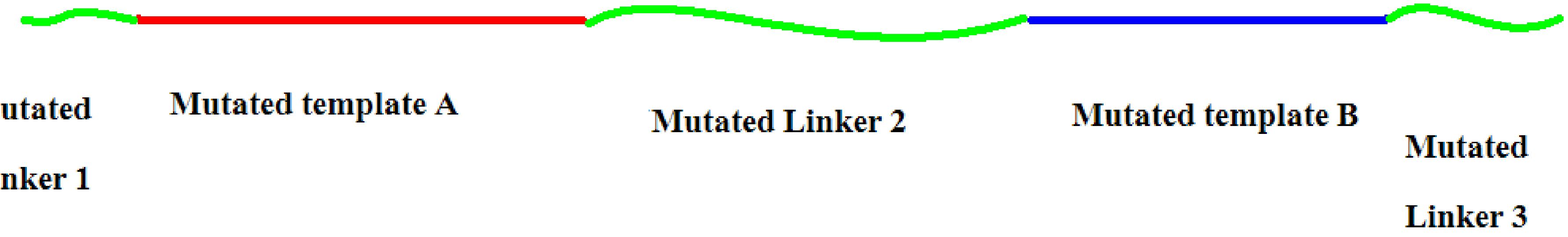
§ Alignments can be improved by including other sources of information, such as predicted secondary structure, or profile-profile, in constructing the alignment

Homology Modeling:

Step Three:

- Link the protein fragments together into one by linkers





Mutated Linker 1

Mutated template A

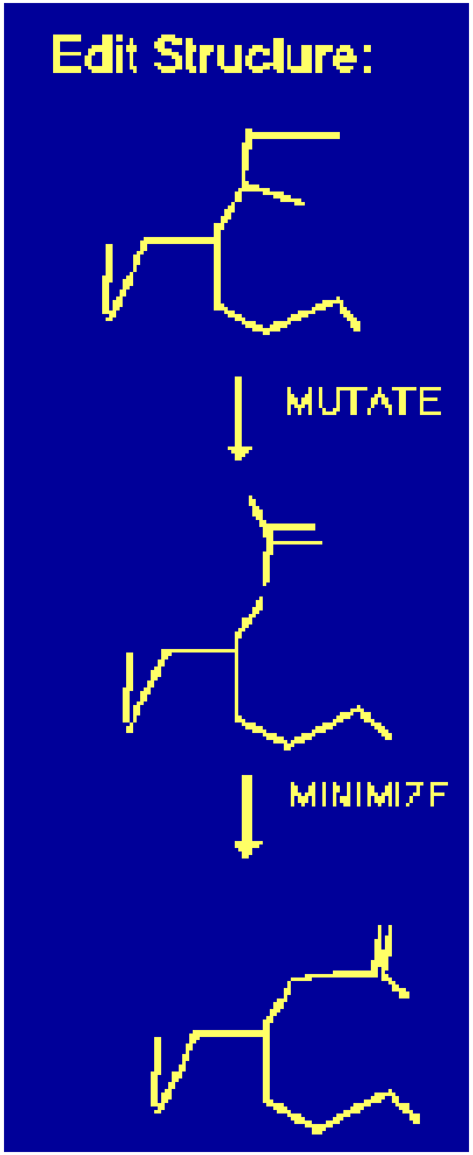
Mutated Linker 2

Mutated template B

Mutated Linker 3

Homology Modeling:

- Step Four: Adding side chains to the main-chain model based on the sequence of your protein:
  - Mutate and add

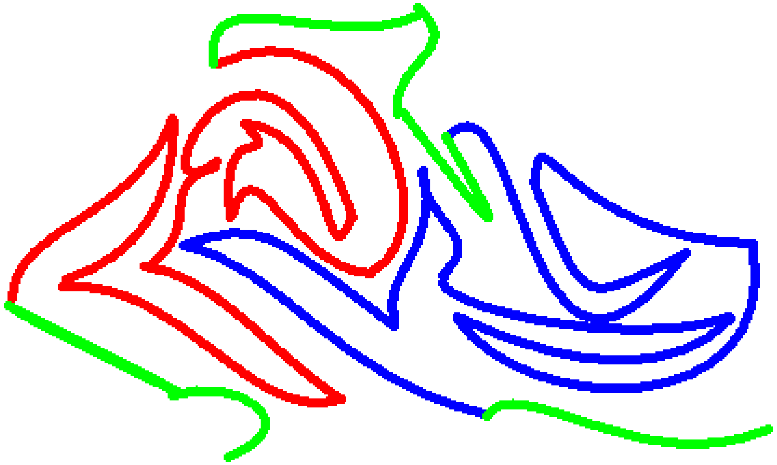


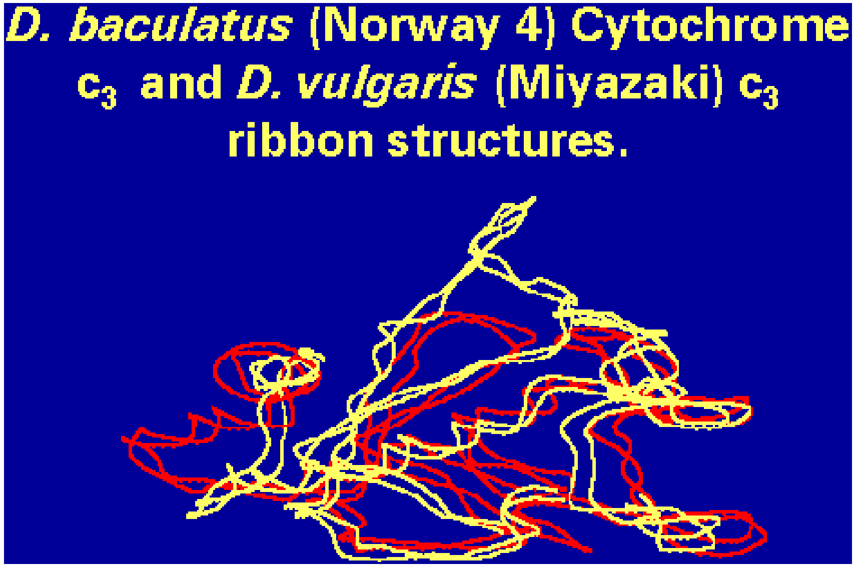
Homology Modeling:

Step Five:

- Minimization and MD of the homology model of your protein

$$H = \sum_{atoms}\frac{p^2}{2m} + \sum_{bond-stretch}\frac{1}{2}k_r(r-r_{eq})^2 + \sum_{bond-angle-bending}\frac{1}{2}k_\theta(\theta-\theta_{eq})^2 +$$
$$\sum_{bond-rotation}\frac{v_n}{2}[1+\cos(n\phi-\gamma)] + \sum_{S-bond}[V_0(1-e^{-a(r-r_0^{'})})^2 - V_0] +$$
$$\sum_{H-bond}[V_0(1-e^{-a(r-r_0^{'})})^2 - V_0] + \sum_{non-bonded}[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\varepsilon_{ij} r_{ij}}]$$

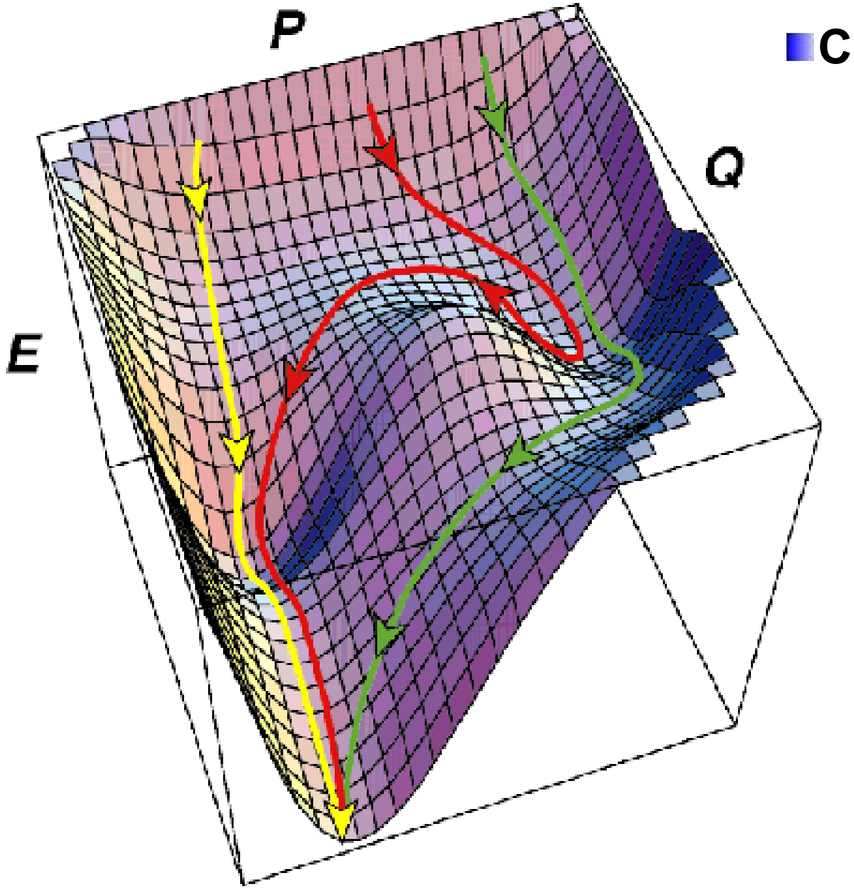




D. baculatus **(Norway 4) Cytochrome c_3 and *D. vulgaris* (Miyazaki) c_3 ribbon structures.**

Search Potential Energy Surface

- We are interested in minimum points on Potential Energy Surface (PES)



- **Conformational search techniques**
  - Energy Minimization
  - Monte Carlo
  - Molecular Dynamics
  - Others: Genetic Algorithm, Simulated Annealing

Homology Modeling:

- Swiss-Model - an automated homology modeling server developed at Glaxo Welcome Experimental Research in Geneva. http://www.expasy.ch/swissmod/

- Closely linked to Swiss-PdbViewer, a tool for viewing and manipulating **protein** structures and models.

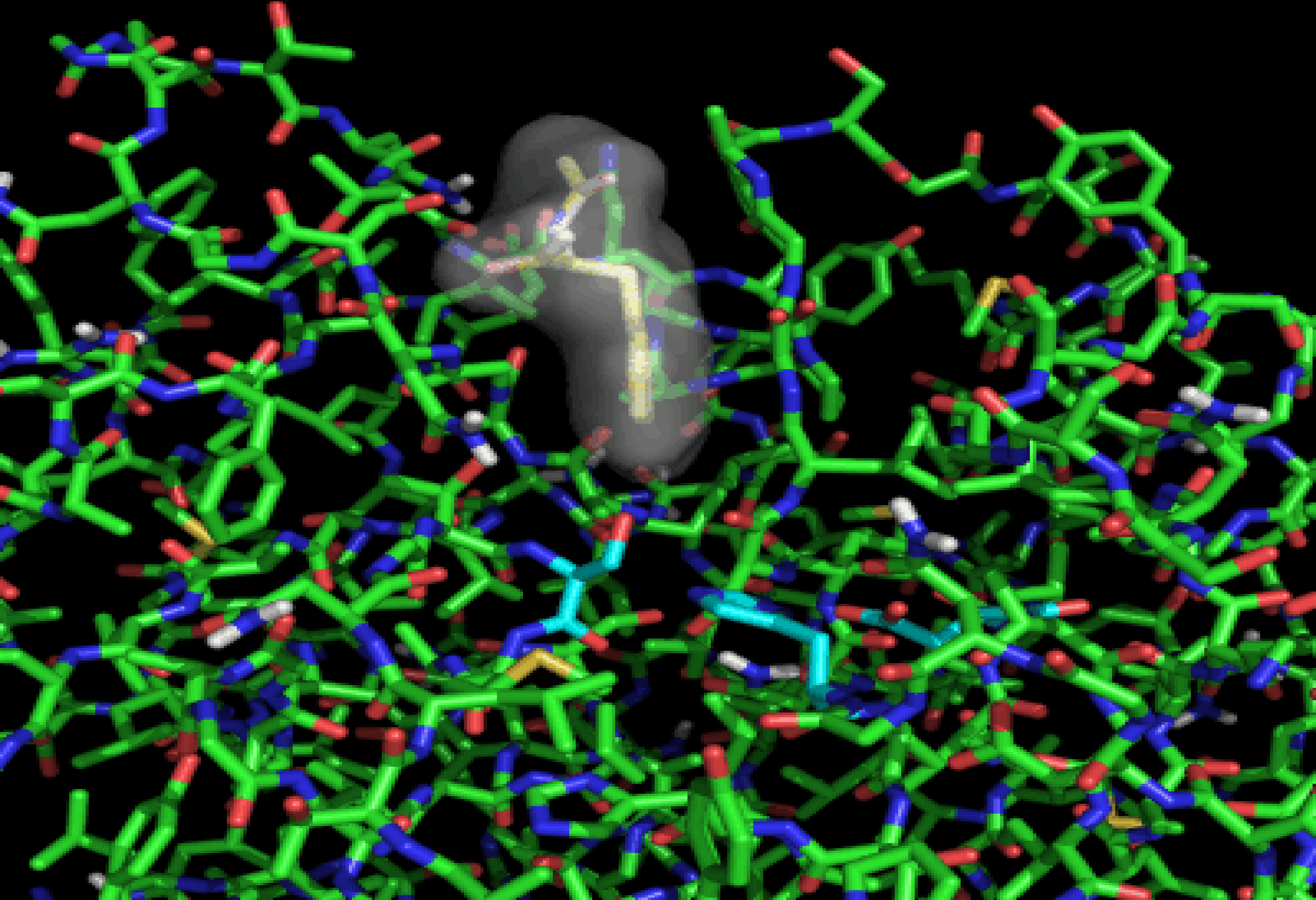
- Likely take 24 hours to get results returned!

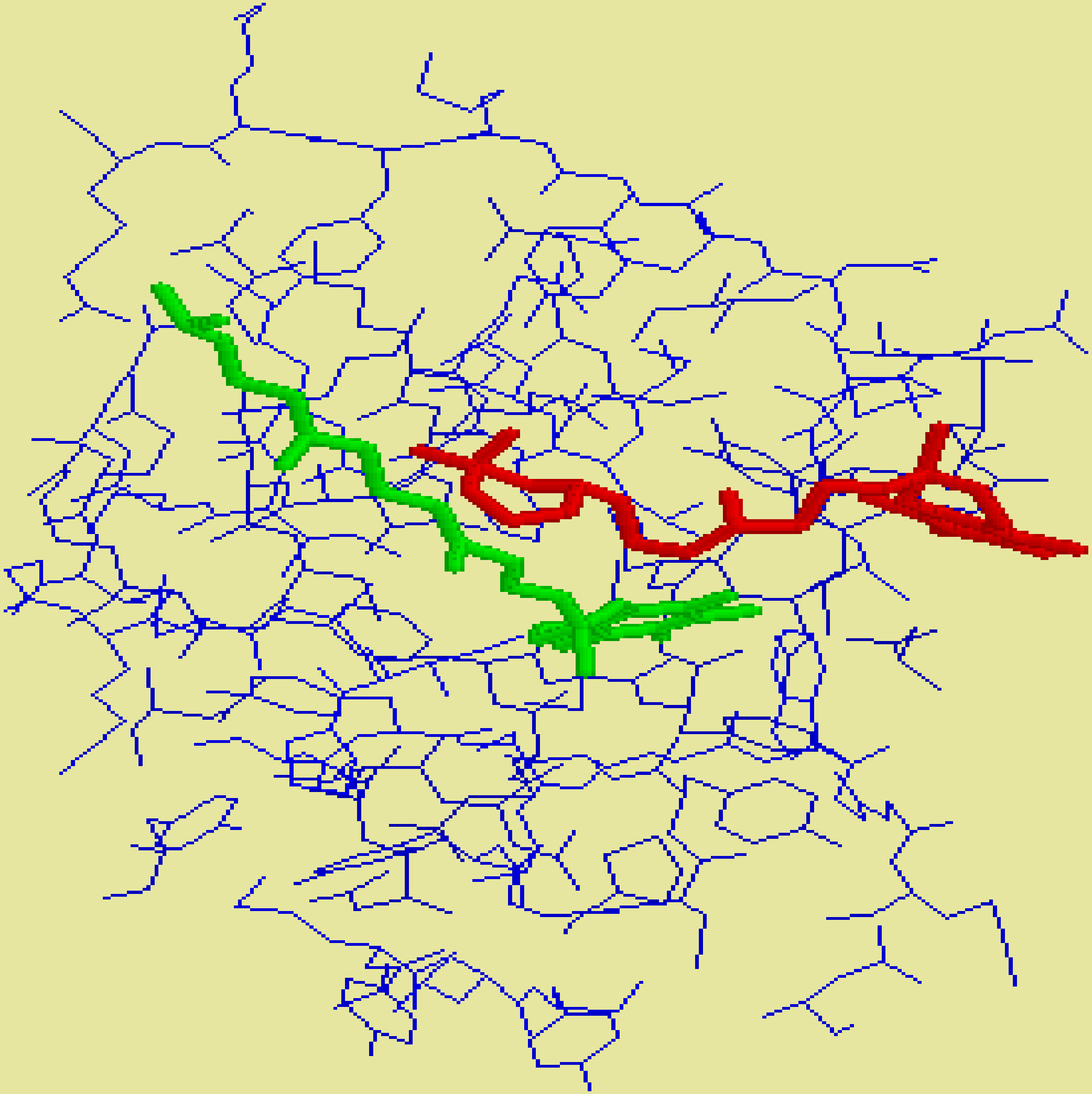
Homology Modeling:

How Swiss-model works?

- 1) Search for suitable templates
- 2) Check sequence identity with target
- 3) Create ProModII jobs
- 4) Generate models with ProModII
- 5) Energy minimization with Gromos96

- First approach mode (regular)
- First approach mode (with user-defined template)
- Optimize mode





课堂作业

- 了解 蛋白结构预测最新方法 Alpha-fold

期末考试: 论文

- 每个人 选一个你感兴趣的 gene/蛋白
- 利用你本节课所学的所有生物信息学数据库、工具、知识， 解读“它”前生今世。
- Deadline： 待公布。