Thousand Faces of Proteins

이주영
고등과학원
2004년 1월 13일

1. **단백질(protein) 왜 중요한가?**
2. **단백질 합성(protein synthesis)과 Central Dogma**
3. **Protein folding problem vs. Traveling salesman problem**
4. **Protein structure (prediction)**

자연과학, 물리

- 힘, 운동, 에너지,
- 우주, 입자,
- 물성,
- 유체의 흐름,
- 자연현상
- 생명현상……

Protein Folding Problem

1. 단백질(protein) 왜 중요한가?

단백질

- DNA, RNA, 탄수화물 (carbohydrate), 지방 (lipid)

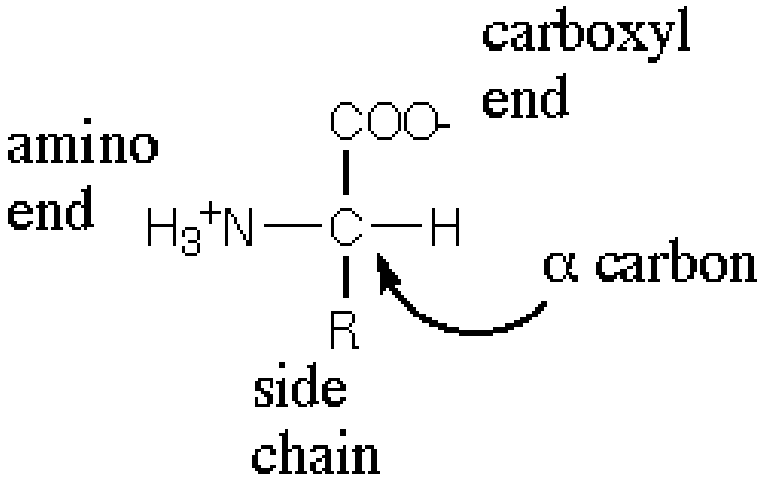
Introduction

- 단백질이란?
  아미노산들이 연결되어 만들어지는 사슬

  cys-gly-val-ala-ala-leu-met

- 생물체에 존재하는 아미노산의 종류: 20가지
- 무수히 많은 조합 가능

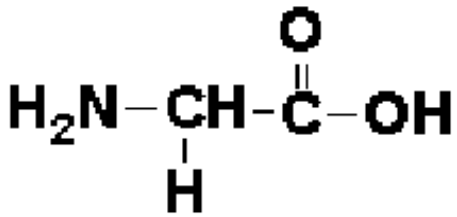
- 아미노산(Amino acid): 단백질의 기본 구성 단위



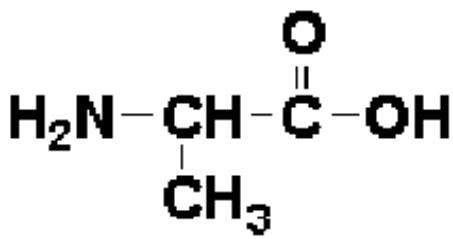
side chain = H, CH_3 ,...

- Side chain에 따라 여러 종류의 아미노산이 만들어진다.
- 생물계에서 주로 발견되는 아미노산은 20가지 종류가 있다.

탈수 결합과 가수 분해



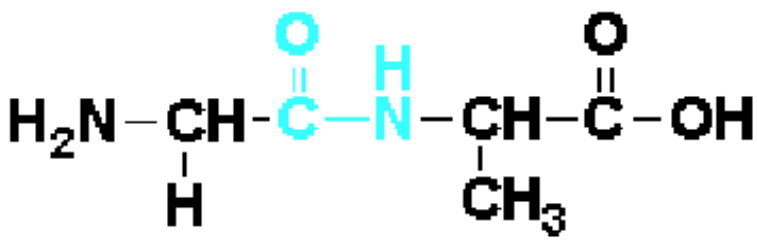
Glycine



Alanine

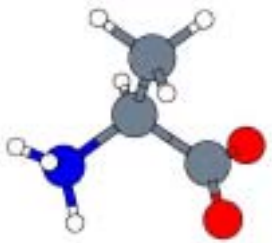
펩티드 결합
(peptide bond)

H_2O

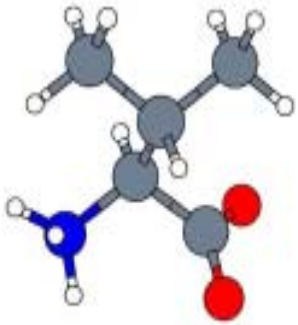


Glycylalanine

물을 싫어하는(hydrophobic) 아미노산

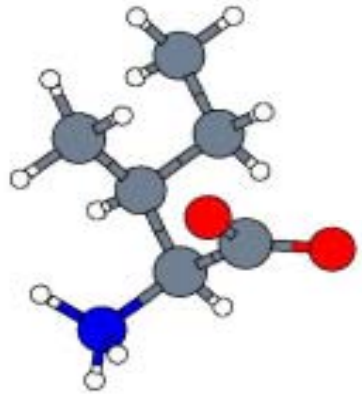


alanine

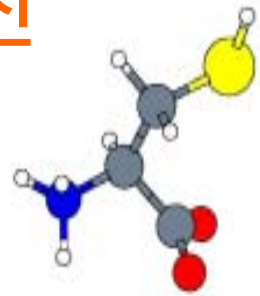


valine

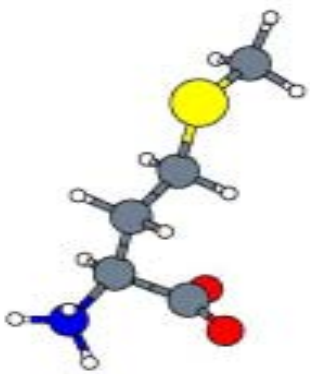
leucine



isoleucine



cysteine



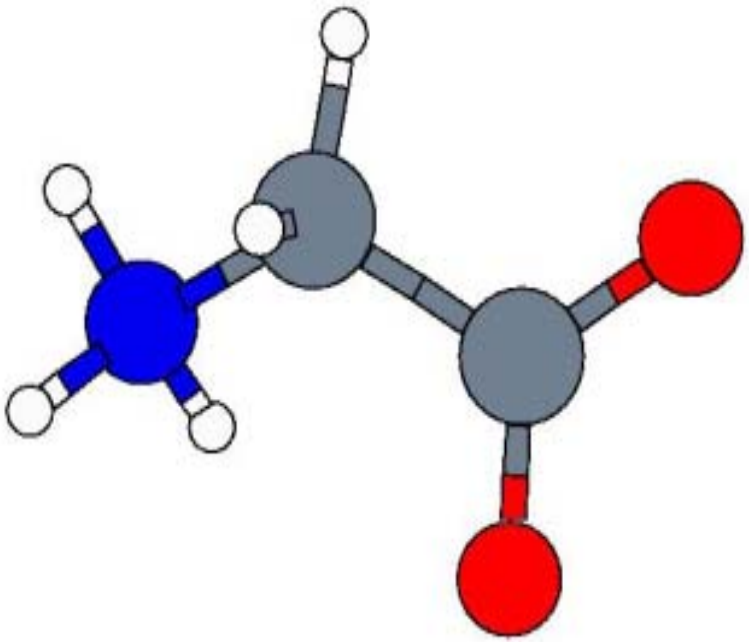
methionine

phenylalanine

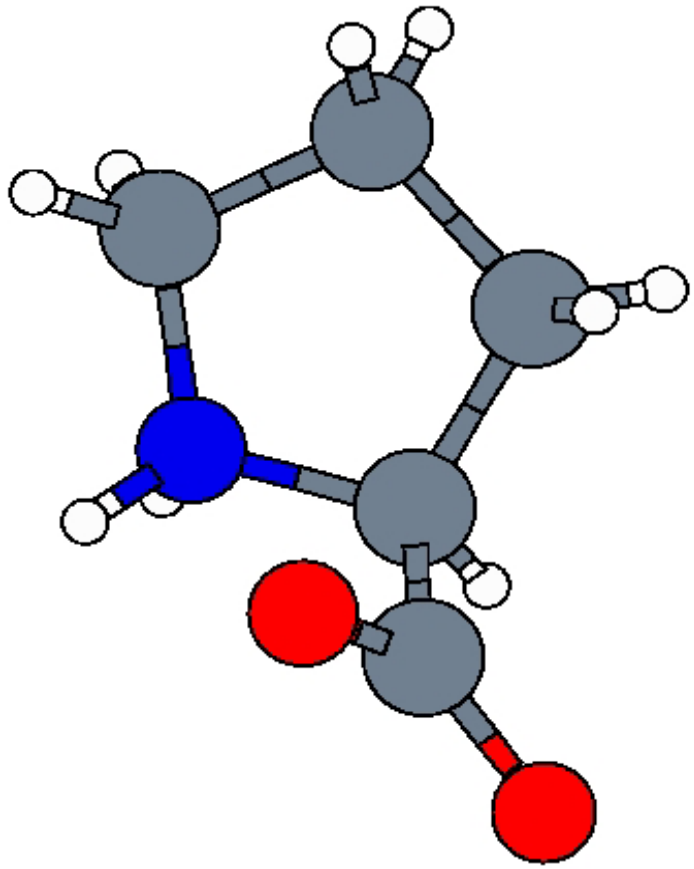
tyrosine

tryptophan

neither hydrophilic nor hydrophobic but rather H_2O-soluble

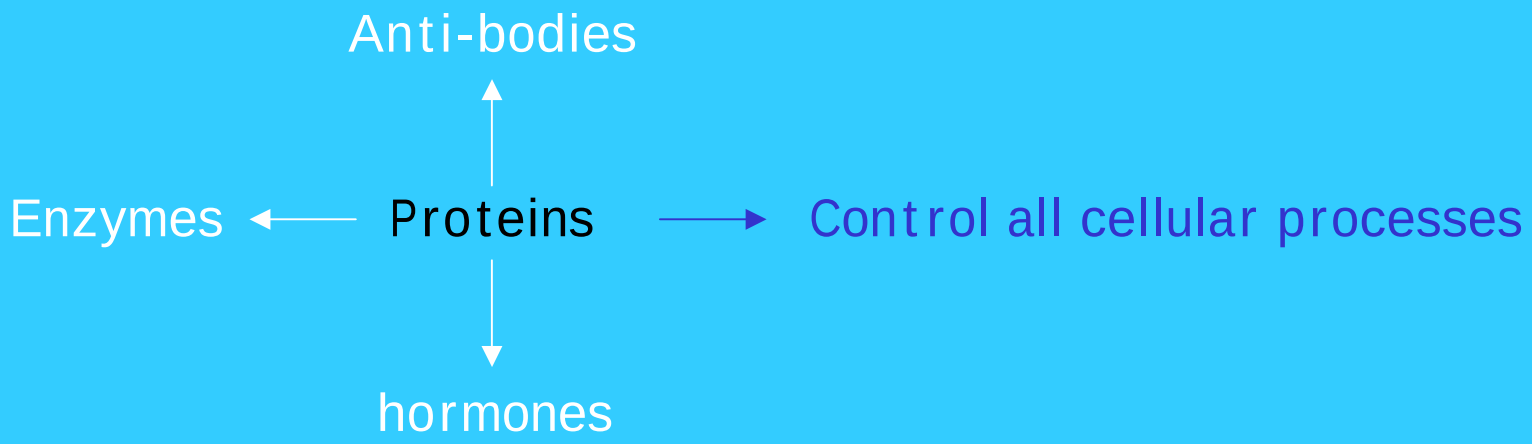


glycine



proline

Why are proteins important ?



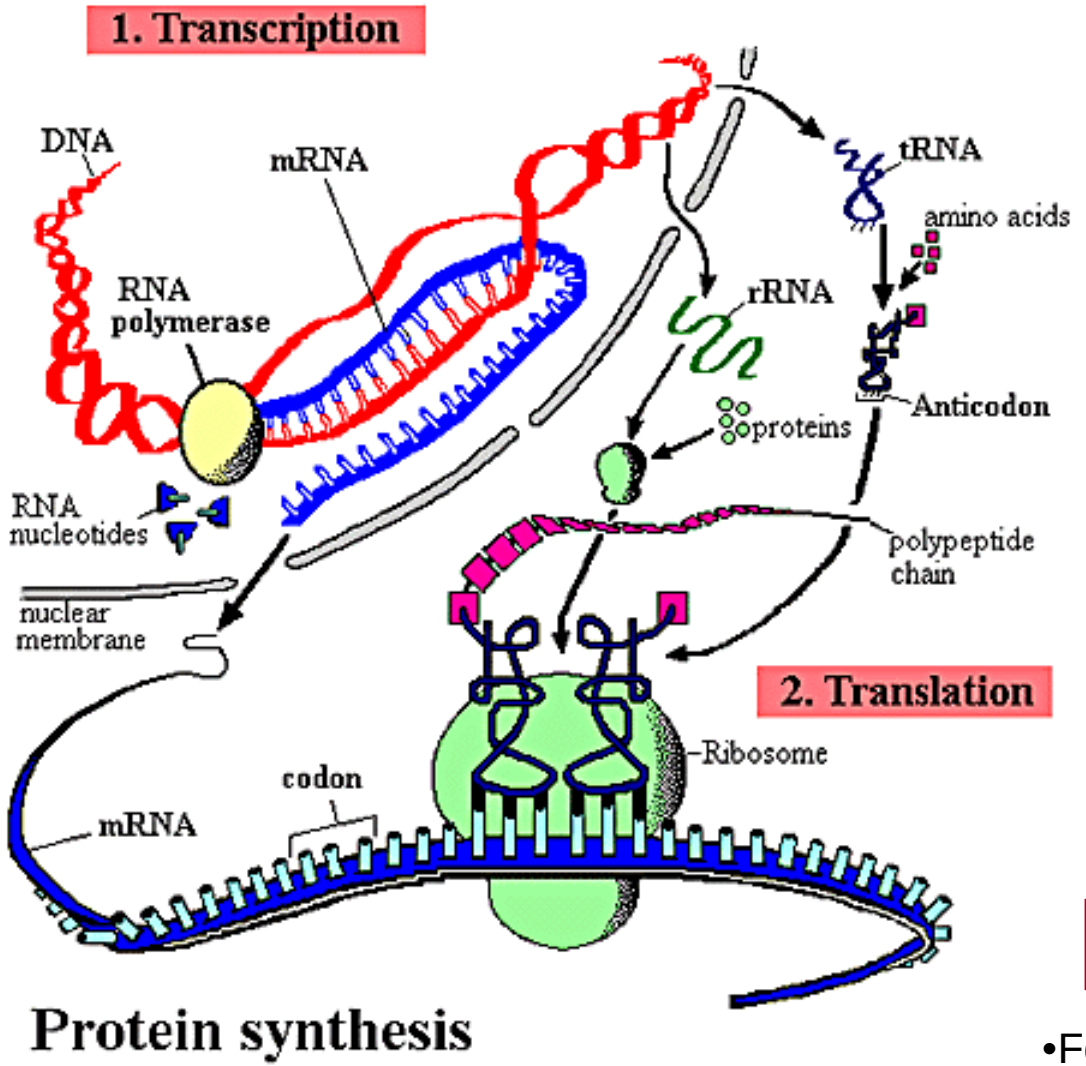
sequence (genome) → (scientific bottleneck: The most challenging problem of this century) → structure (post-genome) → function

(Protein Folding Problem)

Protein Folding Problem

1. 단백질(protein) 왜 중요한가?
2. 단백질 합성(protein synthesis)과 Central Dogma

Central Dogma



Protein synthesis

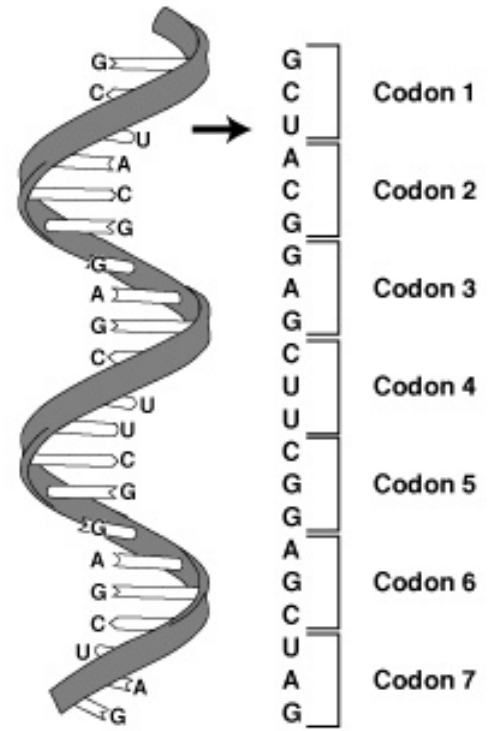
Human hemoglobin

Protein folding problem

- For a given amino acid sequence (of size n), find the native structure of the protein.
- Total # of protein structures: 10^n
- mathematically NOT well defined problem

sequence → structure → function

DNA (RNA), Codon and Amino acids

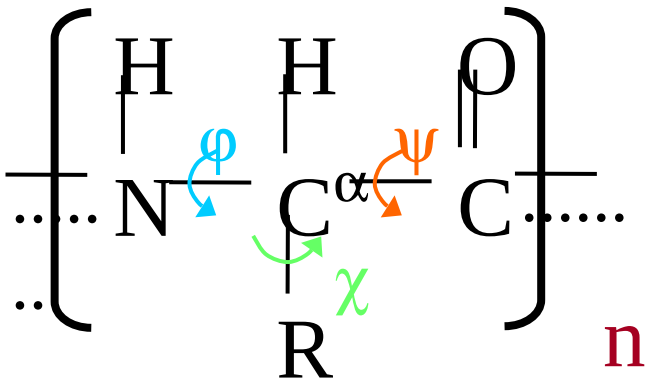


RNA

Ribonucleic acid

1st	2nd T	2nd C	2nd A	2nd G
T	TTT 0.43 Phe TTC 0.57 TTA 0.06 Leu TTG 0.12	TCT 0.18 TCC 0.23 Ser TCA 0.15 TCG 0.06	TAT 0.42 Tyr TAC 0.58 TAA 0.22 TERM TAG 0.17	TGT 0.42 Cys TGC 0.58 TGA 0.61 TERM TGG 1.00 Trp
C	CTT 0.12 CTC 0.20 Leu CTA 0.07 CTG 0.43	CCT 0.29 CCC 0.33 Pro CCA 0.27 CCG 0.11	CAT 0.41 His CAC 0.59 CAA 0.27 Gln CAG 0.73	CGT 0.09 CGC 0.19 Arg CGA 0.10 CGG 0.19
A	ATT 0.35 ATC 0.52 Ile ATA 0.14 ATG 1.00 Met	ACT 0.23 ACC 0.38 Thr ACA 0.27 ACG 0.12	AAT 0.44 Asn AAC 0.56 AAA 0.40 Lys AAG 0.60	AGT 0.14 Ser AGC 0.25 AGA 0.21 Arg AGG 0.22
G	GTT 0.17 GTC 0.25 Val GTA 0.10 GTG 0.48	GCT 0.28 GCC 0.40 Ala GCA 0.22 GCG 0.10	GAT 0.44 Asp GAC 0.56 GAA 0.41 Glu GAG 0.59	GGT 0.18 GGC 0.33 Gly GGA 0.26 GGG 0.23

아미노산 잔기 (Amino Acid Residue)



20 가지 종류

R=H, CH3, CH2-CH2-S-CH3,…

n (# of aa)

- **폴리펩티드(polypeptide): 아미노산 잔기들이 탈수 결합하여 만들어지는 사슬**

- **단백질: 생명체 안에 존재하는 폴리펩티드**

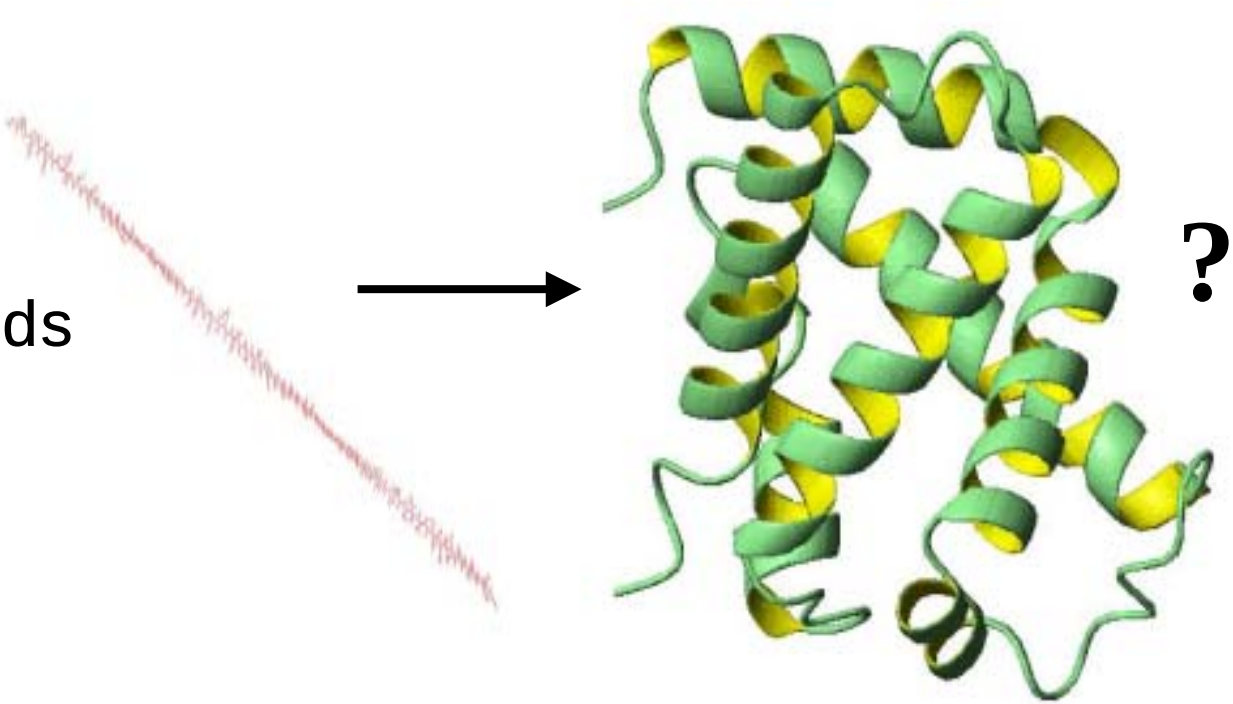
(사람의 경우 약 10^5 종류)

- **단백질은 고유한 3차원 얼개를 갖는다.**

The Protein Folding Problem

1. 단백질 구조 결정

   주어진 아미노산 서열이 있을 때. 그 3차원 얼개(고유 구조:Native structure)는 무엇인가?



141 amino acids

2152 atoms

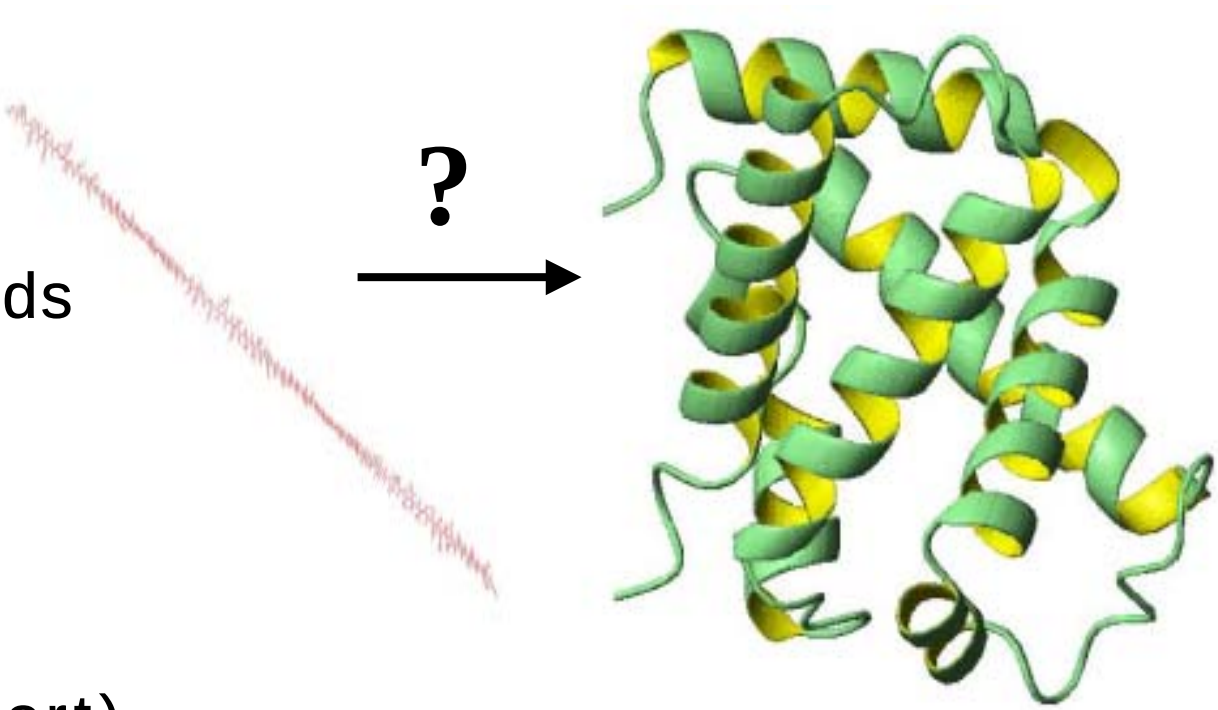
deoxy human hemoglobin

(oxygen transport)

The Protein Folding Problem

2.단백질 접힘 동역학(Protein-Folding Mechanisms)

단백질이 어떤 과정을 통해 고유 구조로 접혀지는가?



141 amino acids

2152 atoms

deoxy human hemoglobin

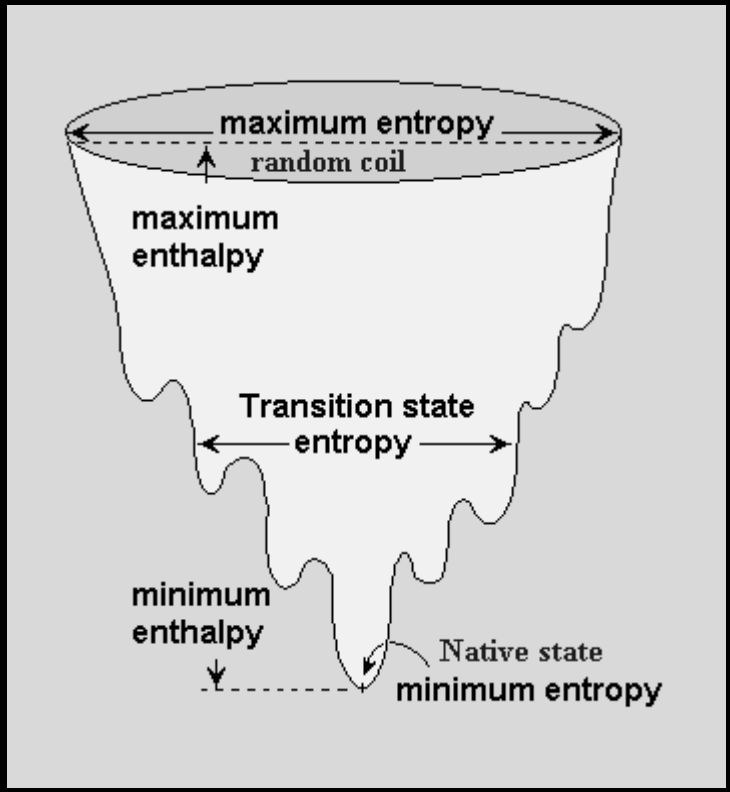
(oxygen transport)

Protein-Folding Mechanisms

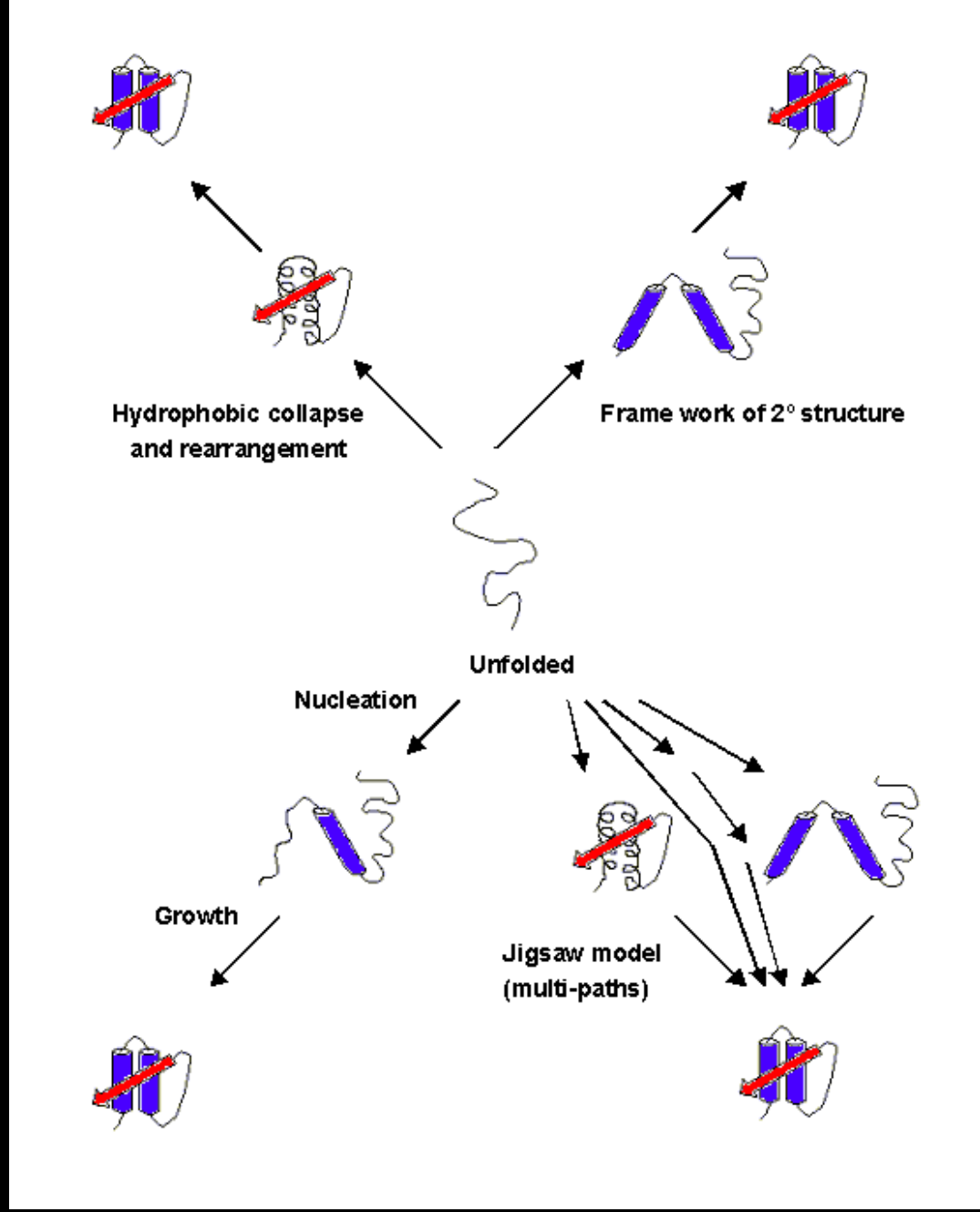
Random search of all conformational space requires an immense amount of time (longer than the age of universe). In vitro refolding normally takes seconds or minutes.
-Levinthal paradox

Ex) for a protein with 100 amino acids, 각각의 아미노산이 helix, sheet, 또는 random 구조중의 한 구조를 갖는다고 가정, 이 사슬이 가질 수 있는 구조의 수는 3^{100} (or about 10^{48}). 만일 각각의 아미노산이 한 구조에서 다른 구조로 변화하는데 10^{14} s^{-1} (작은 분자들에 있어서의 bond rotation 속도)이 걸린다면, random search를 통해 native 구조를 찾는데 걸리는 시간은 10^{34} 초 또는 10^{26} 년 → 우주의 나이보다 더 길다 !

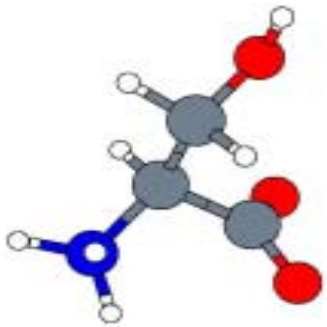
→ **Folding pathway problem: identifying intermediates and constructing folding mechanism**



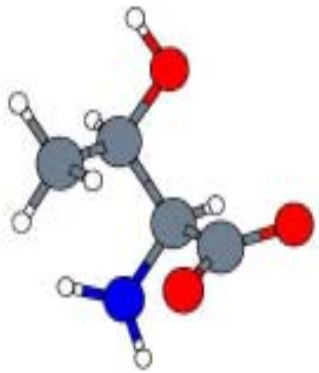
Various models for protein folding



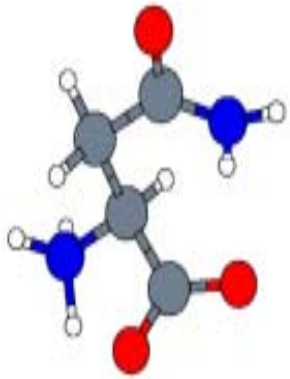
물을 좋아하는(hydrophilic) 아미노산



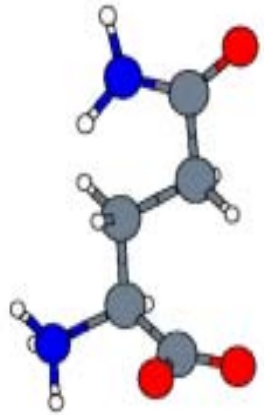
serine



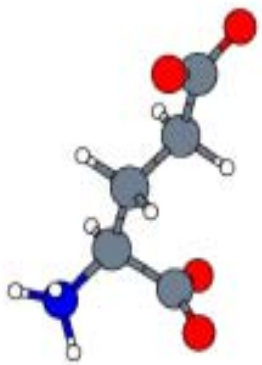
threonine



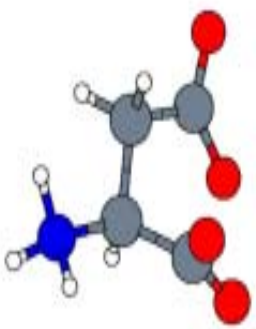
asparagine



glutamine



glutamic acid —



aspartic acid —

lysine +

arginine +

histidine +

•DNA 서열(sequence)

e.g.) ATT-ACG-CAG-CCA-CGG-CGG-ATT

•RNA 서열

UAA-UGC-GUC-GGU-GCC-GCC-UAA

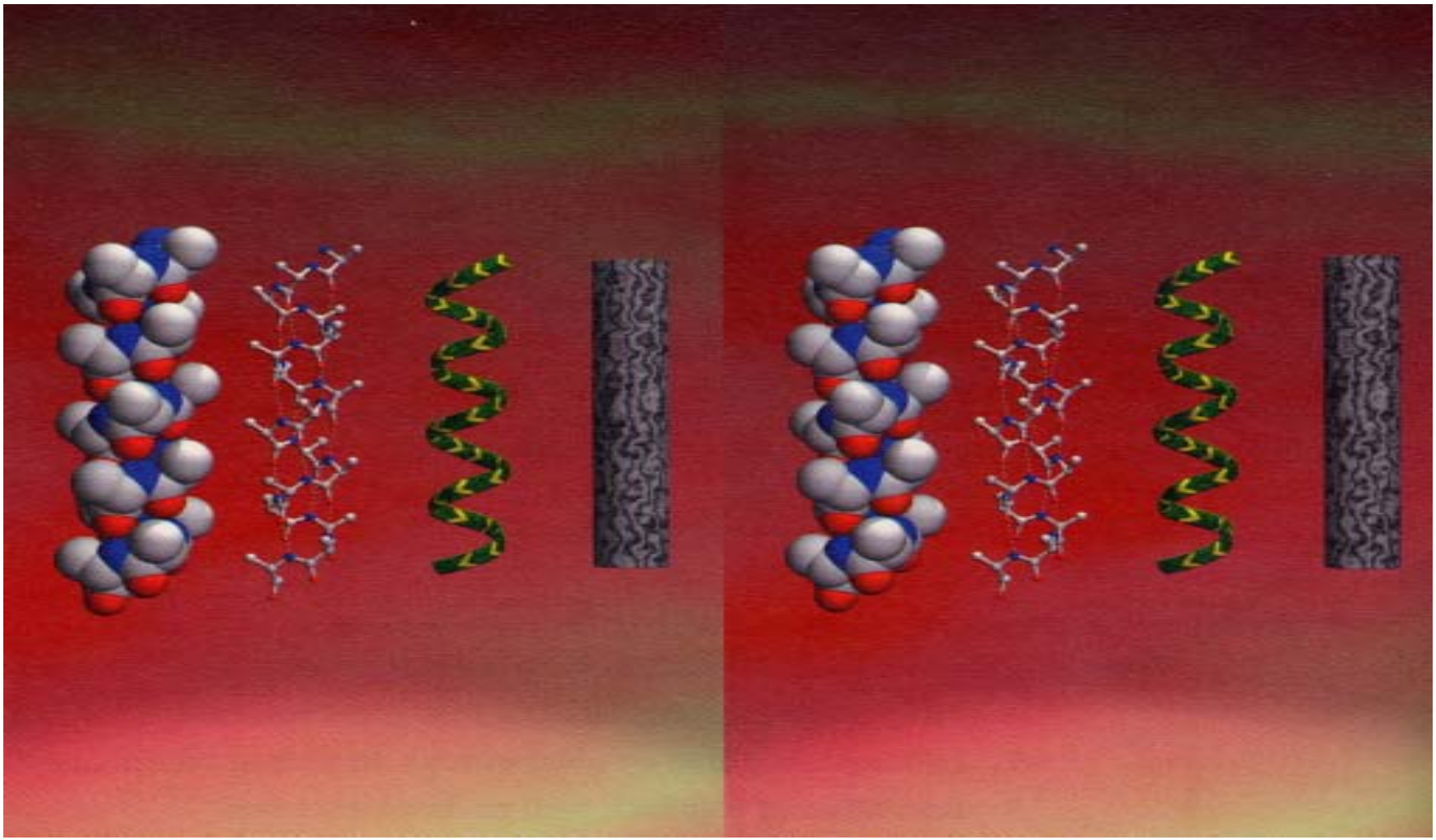
•아미노산 서열(Amino acid sequence)

cys-gly-val-ala-ala

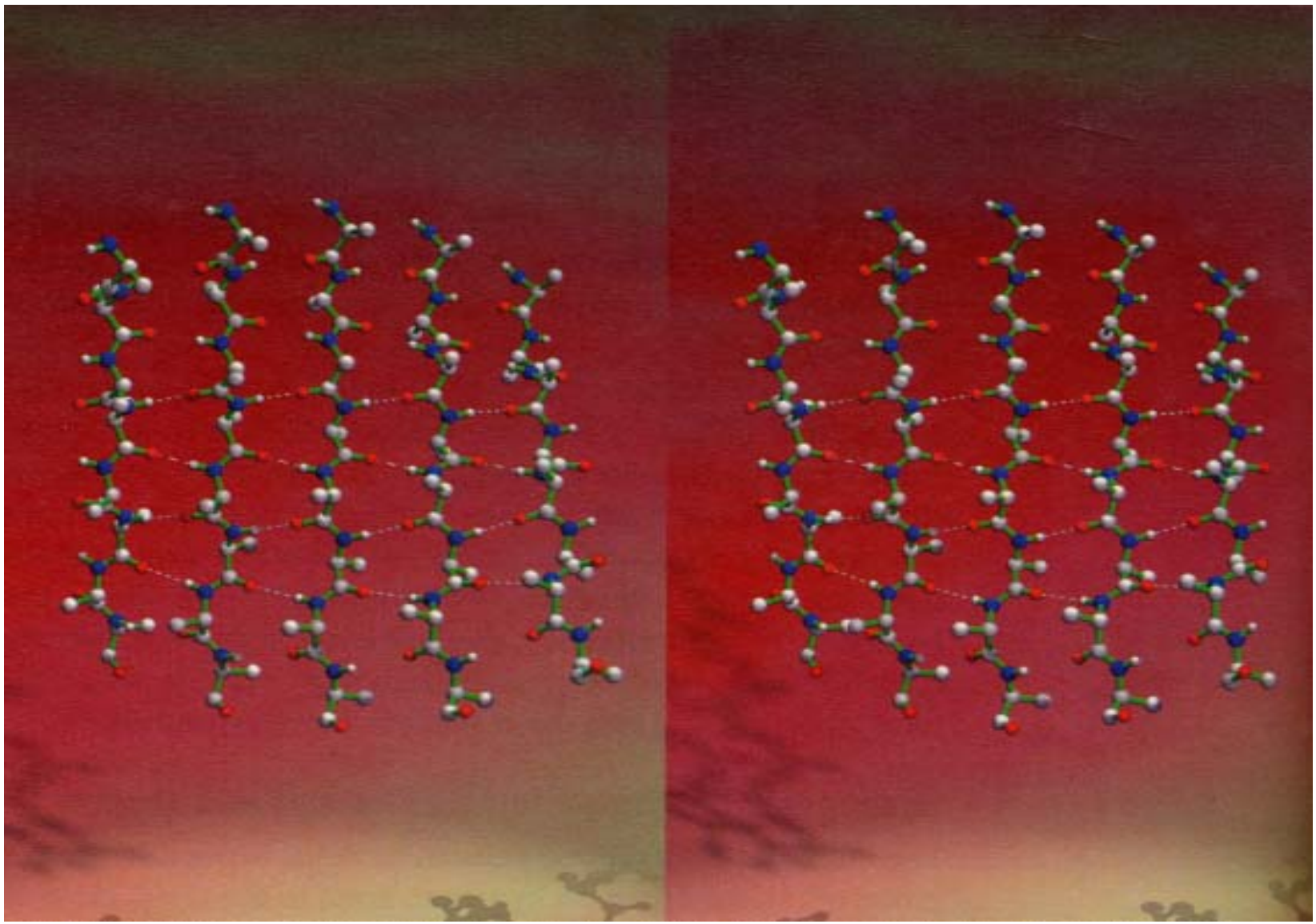
→ 1차 구조(Primary Structure)

- secondary structure: local conformation of backbones

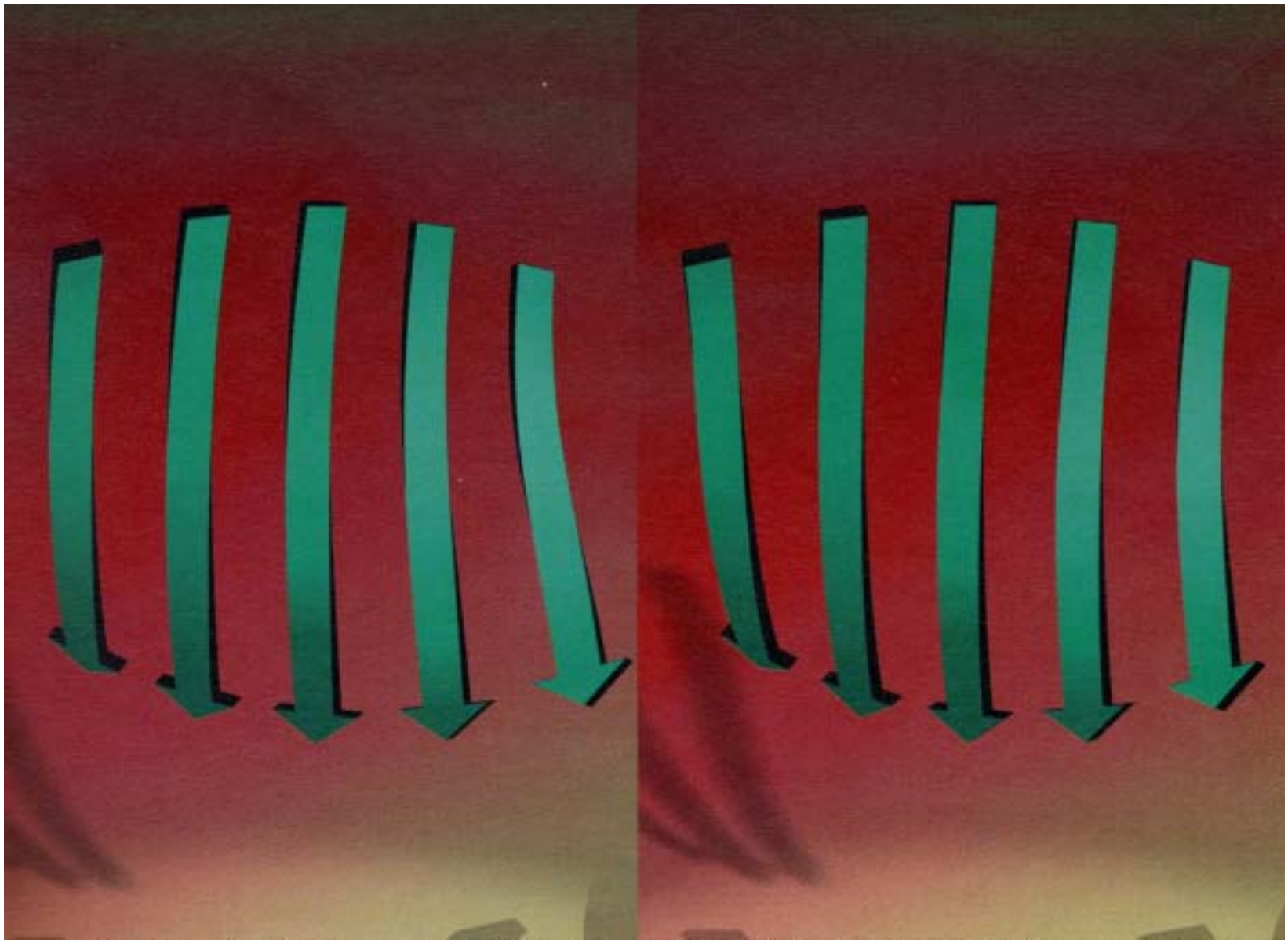
e.g.) α helix



e.g.) β sheet



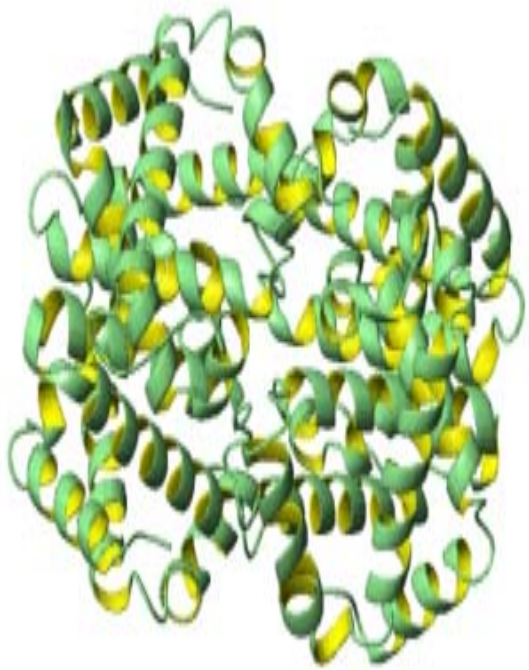
β sheet (continued)



•3차 구조(Tertiary Structure):
overall topology of a folded protein

crambin (46 aa)

•4차 구조(Quaternary Structure)



deoxy human hemoglobin

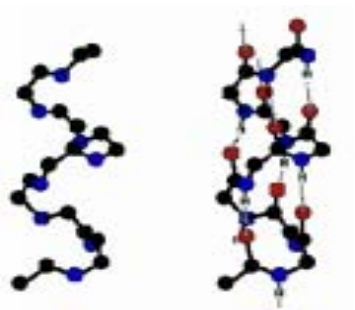
(oxygen transport)

4 proteins

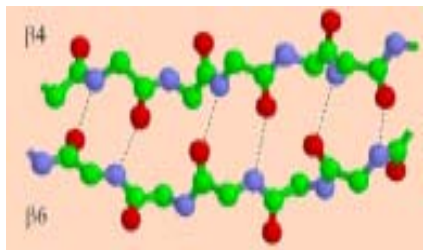
141-146-141-146 aa

- 1차 구조(Primary Structure)
  1D sequence of amino acids
  e.g.)cys-gly-val-ala-ala

- 2차 구조(Secondary Structure)



α helix



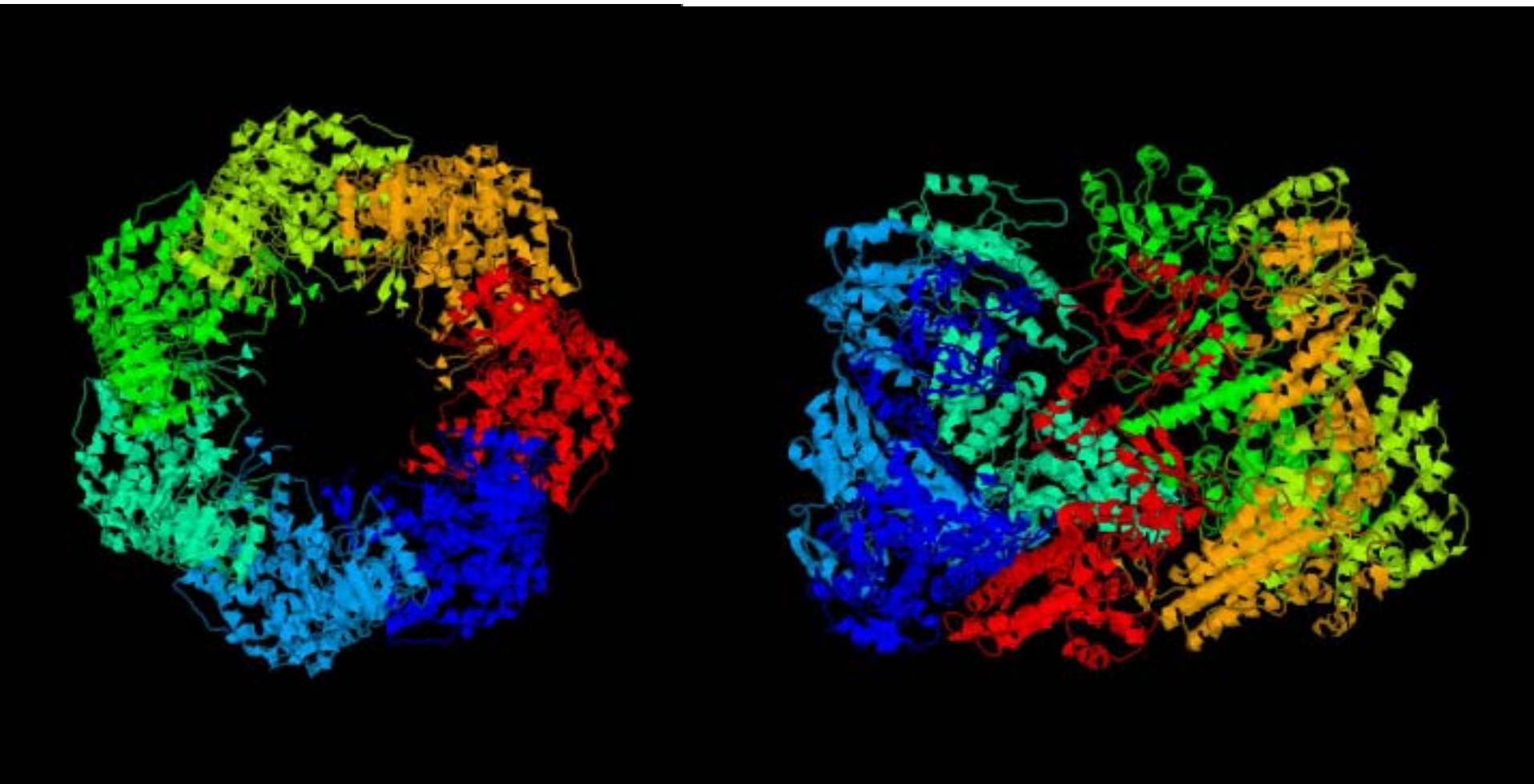
β sheet

- 3차 구조(Tertiary Structure)
  3D arrangement

GroEL

top view

side view



1. Protein Structure Determination

A. 실험적인 방법:

X-ray crystallography,

NMR

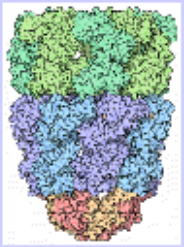
J. Kendrew (1957): myoglobin

M. Perutz (1959): hemoglobin

DEPOSIT data
DOWNLOAD files
browse LINKS
BETA TEST new features

Current Holdings

18488 Structures
Last Update: 20-Aug-2002
PDB Statistics



Molecule of the Month:
Chaperones

The Protein Data Bank (PDB) is operated by Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the National Institute of Standards and Technology -- three members of the Research Collaboratory for Structural Bioinformatics (RCSB). The PDB is supported by funds from the National Science Foundation, the Department of Energy, and two units of the National Institutes of Health: the National Institute of General Medical Sciences and the National Library of Medicine.

PDB
PROTEIN DATA BANK

Welcome to the PDB, the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data.

RCSB Home Contact Us Help

Did you find what you wanted?

Search the Archive

Enter a PDB ID or keyword

Query Tutorial

Find a structure

query by PDB id only match exact word
remove sequence homologues

SearchLite keyword search form with examples
SearchFields customizable search form
Status Search find entries awaiting release

News

Complete News
Newsletter
pdb-l Mailing List
Subscribe

20-Aug-2002
New mmCIF and Data Processing Software Available RCSB-developed programs for mmCIF and data processing -- including MAXIT, PDB_EXTRACT, and an mmCIF database loader -- are now available from the PDB's software page for download... [MORE...]

PDB Mirrors

Please bookmark a mirror site

San Diego Supercomputer Center*
Rutgers University*
National Institute of Standards and Technology*
Cambridge Crystallographic Data Centre, UK
National University of Singapore
Osaka University, Japan
Universidade Federal de Minas Gerais, Brazil

OTHER SITES

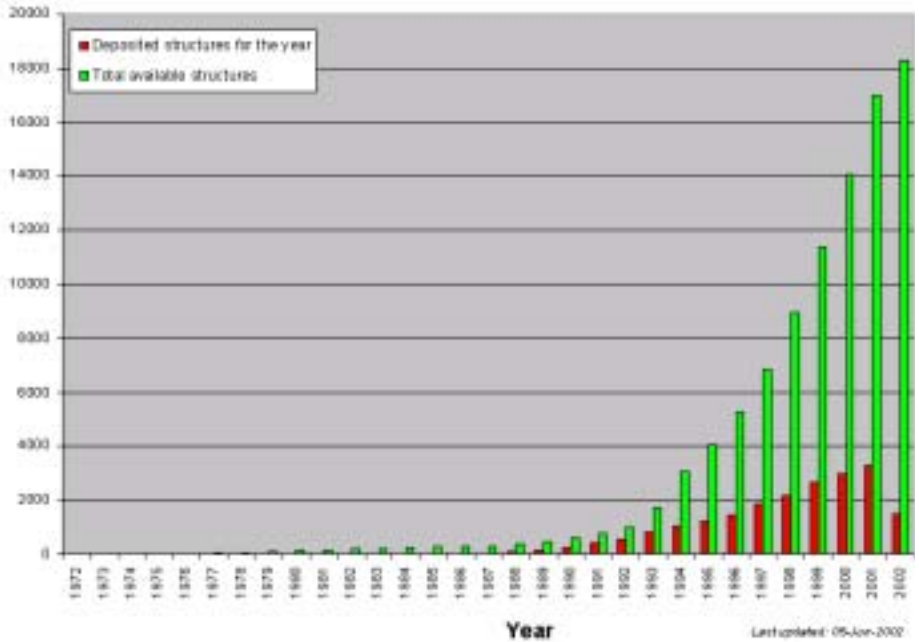
**RCSB partner*

In citing the PDB please refer to:

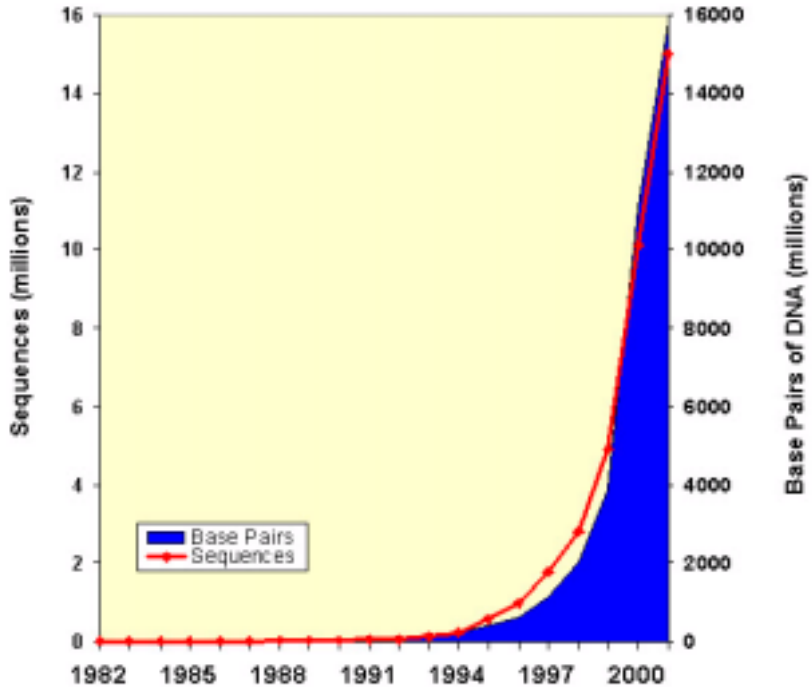
H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The Protein Data Bank. *Nucleic Acids Research*, **28** pp. 235-242 (2000)

ABOUT PDB | DATA UNIFORMITY | RECENT FEATURES | USER GUIDES | FILE FORMATS | EDUCATION | STRUCTURAL GENOMICS | PUBLICATIONS | SOFTWARE

PDB Content Growth



Growth of GenBank



B. 이론적인 방법 (단백질 구조 예측)

knowledge-based methods

이미 실험으로 구한, 다른 단백질들의 구조를 이용하는 방법.(**Protein Data Bank(PDB)**에 약 **20000**종의 단백질 구조가 있음)

physics-based methods (ab initio prediction)

물리적인 원리의 이해를 통해 구조를 예측

Knowledge-based method

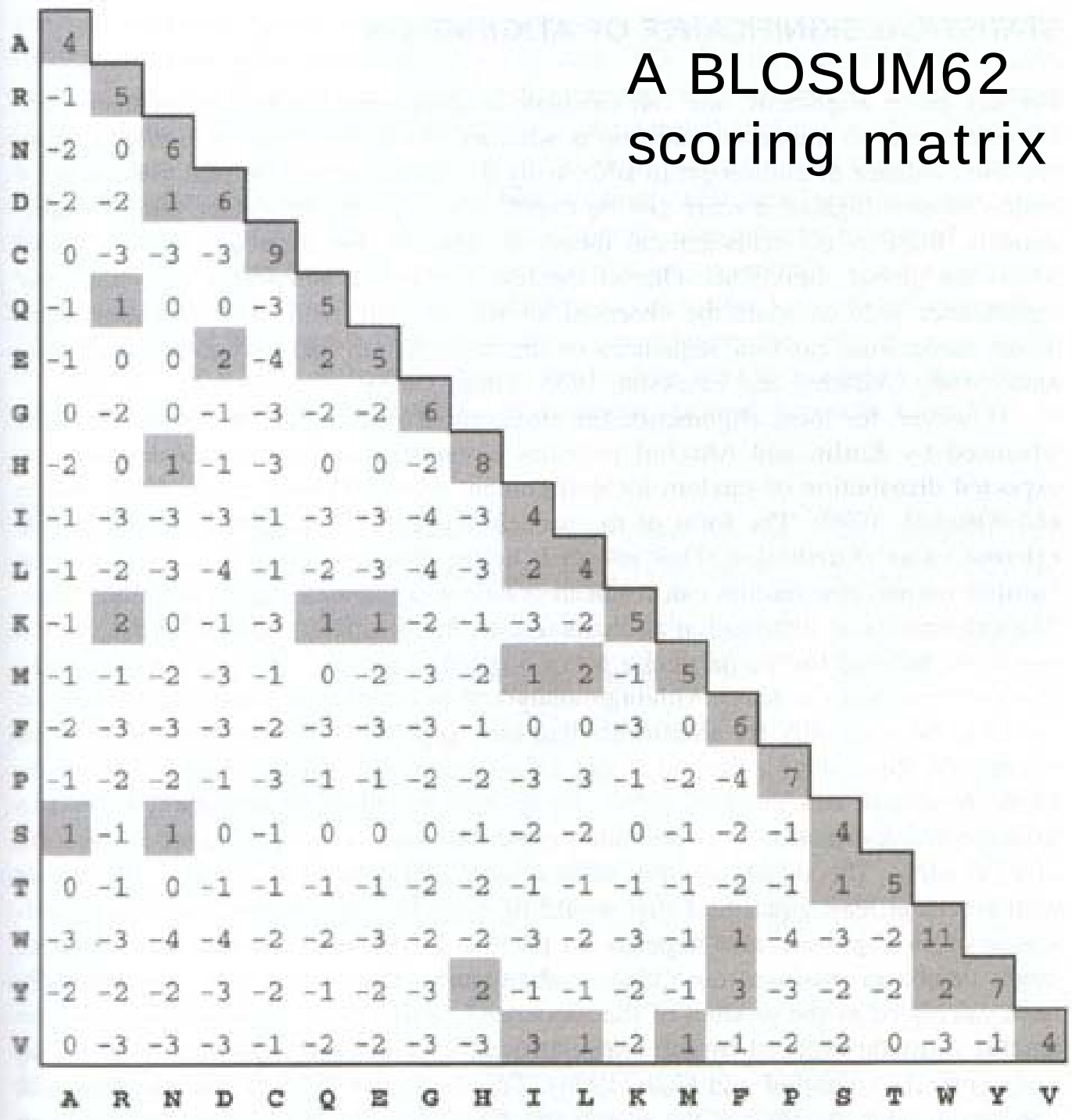
- Comparative modeling (Homology modeling) :
  - PDB 에 **비슷한 서열**이 있다면 구조도 또한 비슷할 것이라 가정한다.
  - 비슷한 서열의 구조를 가져다 쓴다.
  - PDB에 상당히 비슷한 서열이 존재할 때 유용하다.(high sequence similarity)

- Fold recognition (Threading) :
  - PDB 에 존재하는 구조들을 분류한다.
  - 새로운 서열의 구조도 그 중 하나일 것이라 가정한다.
  - 에너지 등을 복합적으로 고려하여 가장 적합한 구조를 가져다 쓴다.
  - PDB에 어느 정도 비슷한 서열이 존재할 때 유용하다.(medium sequence similarity)

Sequences and Sequence allignment

- Two main kind of sequences
  - Sequence of base pairs in DNA molecules
    - (A+T+C+G)*
  - Sequence of aminoacids in a protein molecule
    - A(C+D+E+F+G+H+I+K+L+M+N+P+Q+R+S+T+V+W+X+Y)*Z
- Two main kind of sequence allignment
  - Global alignment
    - LGPSSKQT GKG S - -RIWDN
    - | | | | | | |
    - LN –ITKSA GKGAIMRLGDA
  - Local alignment
    - ----------TGKG----------
    - | | |
    - ----------AGKG----------

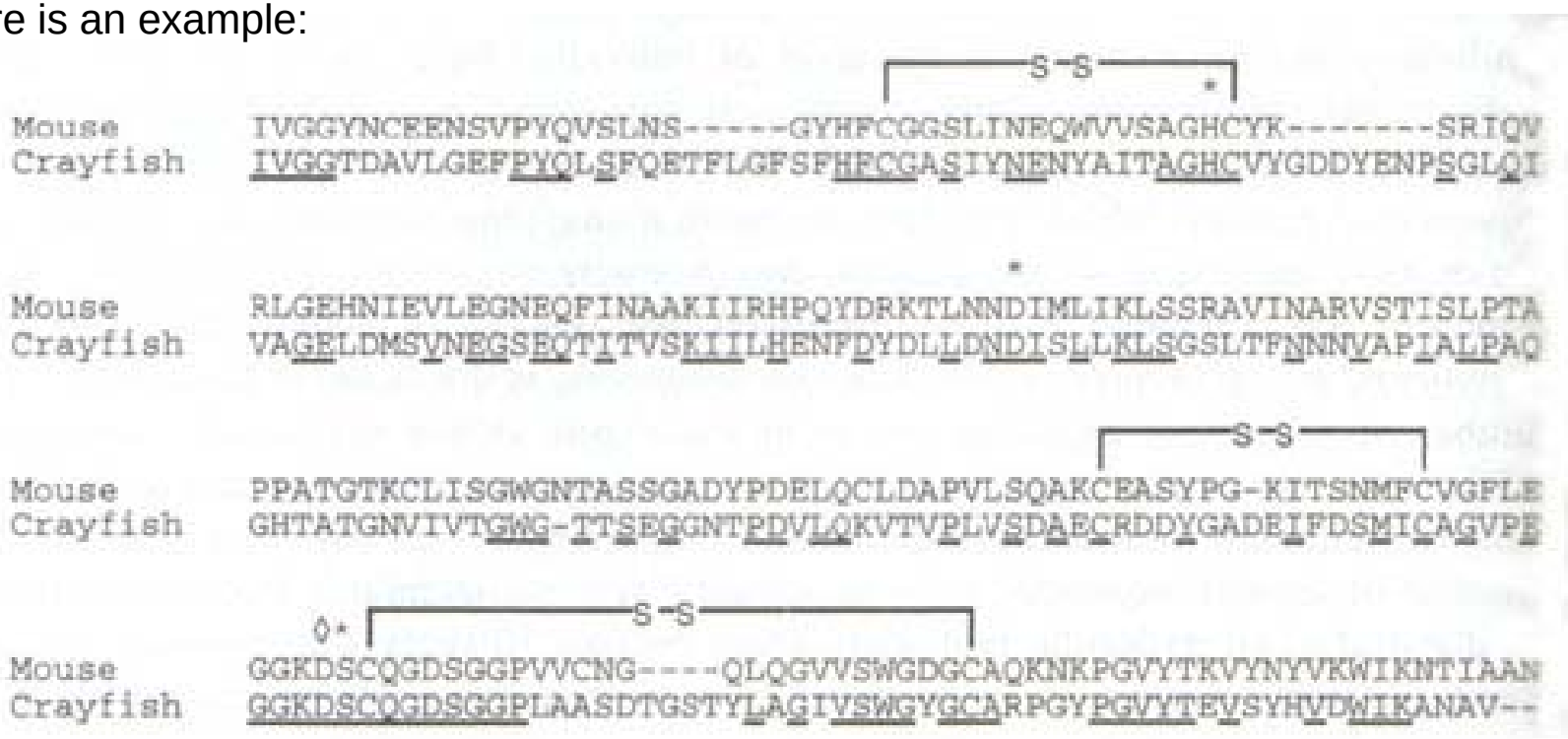
A BLOSUM62 scoring matrix



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Basis of Basic Sequence Alignment

The objective of sequence alignment is to determine if 2 sequences are sufficiently similar to declare them homologous.
Here is an example:



The above illustration shows the alignment of trypsin proteins of mouse (SWISS-PROT P07146) abd crayfish (SWISS-PROT P00765) Identical residues are underlined. Indicated above the alignments are 3 disulfide bonds (-S-S-) with participating cysteine residues conserved, amino acids side chains involved in the charge relay system (asterisk), and active side residue governing substrate specificity (diamond).

Ab-initio(energy-based,Physics-based) method

- 통계역학 원리에 바탕을 둔다.
- PDB에 비슷한 서열이 없을 때 유용하다. (no sequence similarity)
- 접힘 동역학도 연구할 수 있다.

- There was a student who knew nothing about quantum mechanics
- This poor student took a “quantum mechanics” course.
- He had to take a take-home exam: Not to drop out of the class, he has two options to choose:

1. Examine last 40 year’s problem set **with answers**. → Homology modeling / Threading.

2. Try to understand the problems and write down his own answers. → Ab initio (de novo, Energy-based,…).

Physics-based method ←통계역학 원리:

1차원적인 아미노산 서열이 3차원 구조를 결정한다. (Anfinsen, 1973)

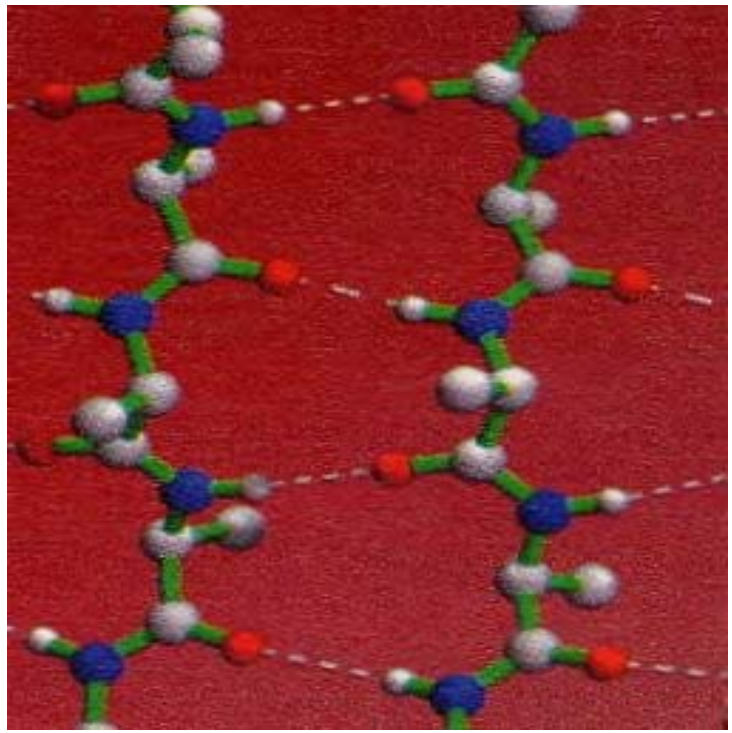
→단백질 접힘

통계역학의 관점에서 볼 때, 고유 구조란 자유 에너지 함수 F가 최저값을 갖는 구조이다.

F=E-TS (E: energy, T: temperature,
S: entropy)

에너지

아미노산 잔기들 간의 온갖 상호작용을 알아야 한다.
(Electrostatic interaction between polar atoms,
Van der Waals force,
Disulfide bridge,
Etc.)



Various Potential Energy Functions

- All-atom Consideration
- Coarse-grained Function

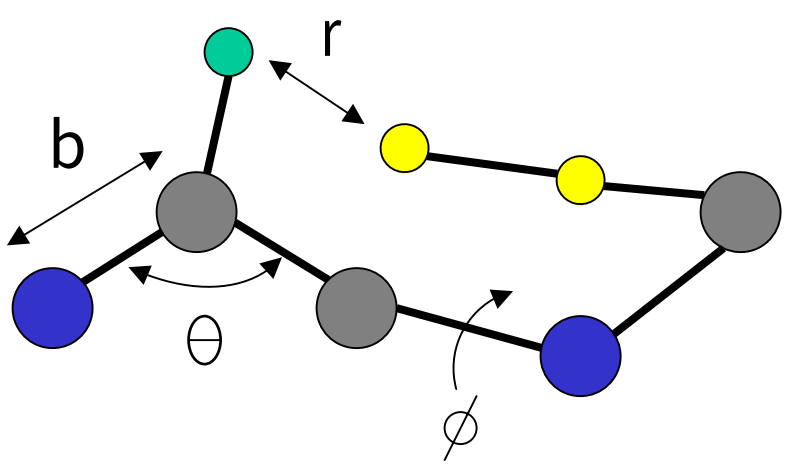
  Atom-centered, Fixed charge…
- Contacts only : Scoring Function

Potential energy functions

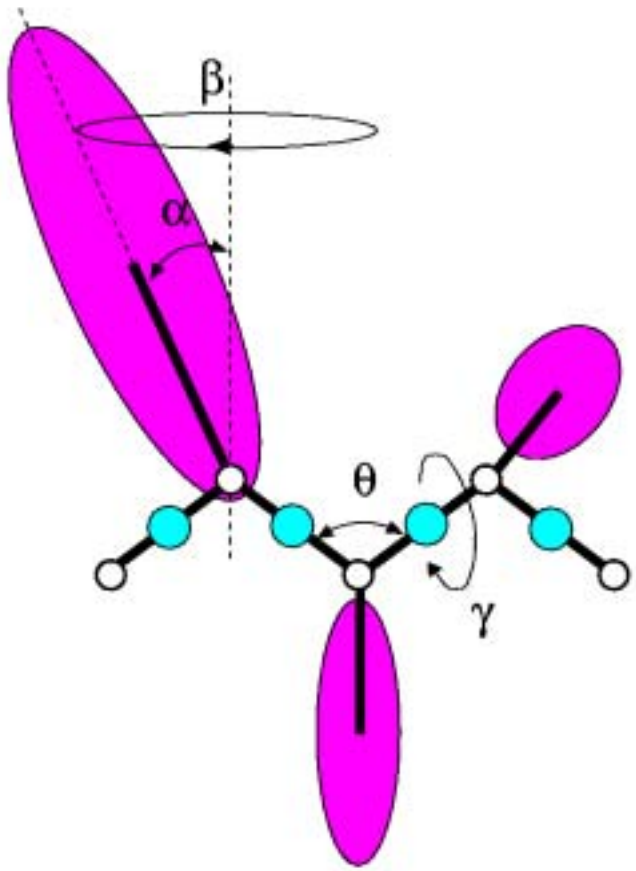
1. All atom, off-lattice potentials
   - ECEPP, AMBER, CHARMM, and more
   - E terms: vibrational, torsional, non-bonded, electrostatic

$$E = \frac{1}{2}\sum_{\text{bonds}} k_b (b - b_{\text{eq}})^2 + \frac{1}{2}\sum_{\substack{\text{bonds}\\ \text{angles}}} k_\theta (\theta - \theta_{\text{eq}})^2 +$$

$$\sum_{\substack{\text{dihedral}\\ \text{angles}}} k_\phi \cos(n\phi - \delta) + \sum_{ij} \left(\frac{A_{ij}}{r_{ij}^6} + \frac{B_{ij}}{r_{ij}^{12}} + \frac{C_{ij}}{r_{ij}^{10}} + \frac{q_i q_j}{D r_{ij}}\right)$$



2. Coarse grained potentials:
   e.g) UNRES(United-residue potential)
   - Fixed bond (virtual bond) lengths
   - Two interacting centers per residue
   - Can treat larger molecules in reasonable CPU time
   - Off-lattice model

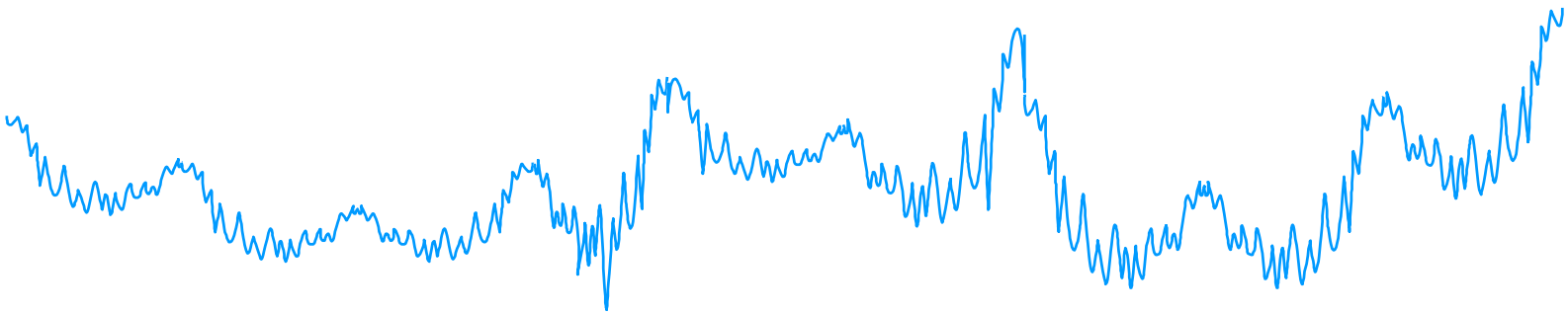


Definition of degrees of freedom for the UNRES representation of a polypeptide chain

Side-chains are represented as ellipsoids (Gay-Berne potential)

Interaction centers are marked in colors

최적화(Global optimization)



- 주어진 함수의 최소값을 주는 점(global minimum)을 찾아야 한다.

최적화(Global optimization)

- **Simulated Annealing (SA)**
- **Genetic Algorithms (GA)**
- **Hopfield Neural Network**
- **Monte Carlo with Minimization**

.

.

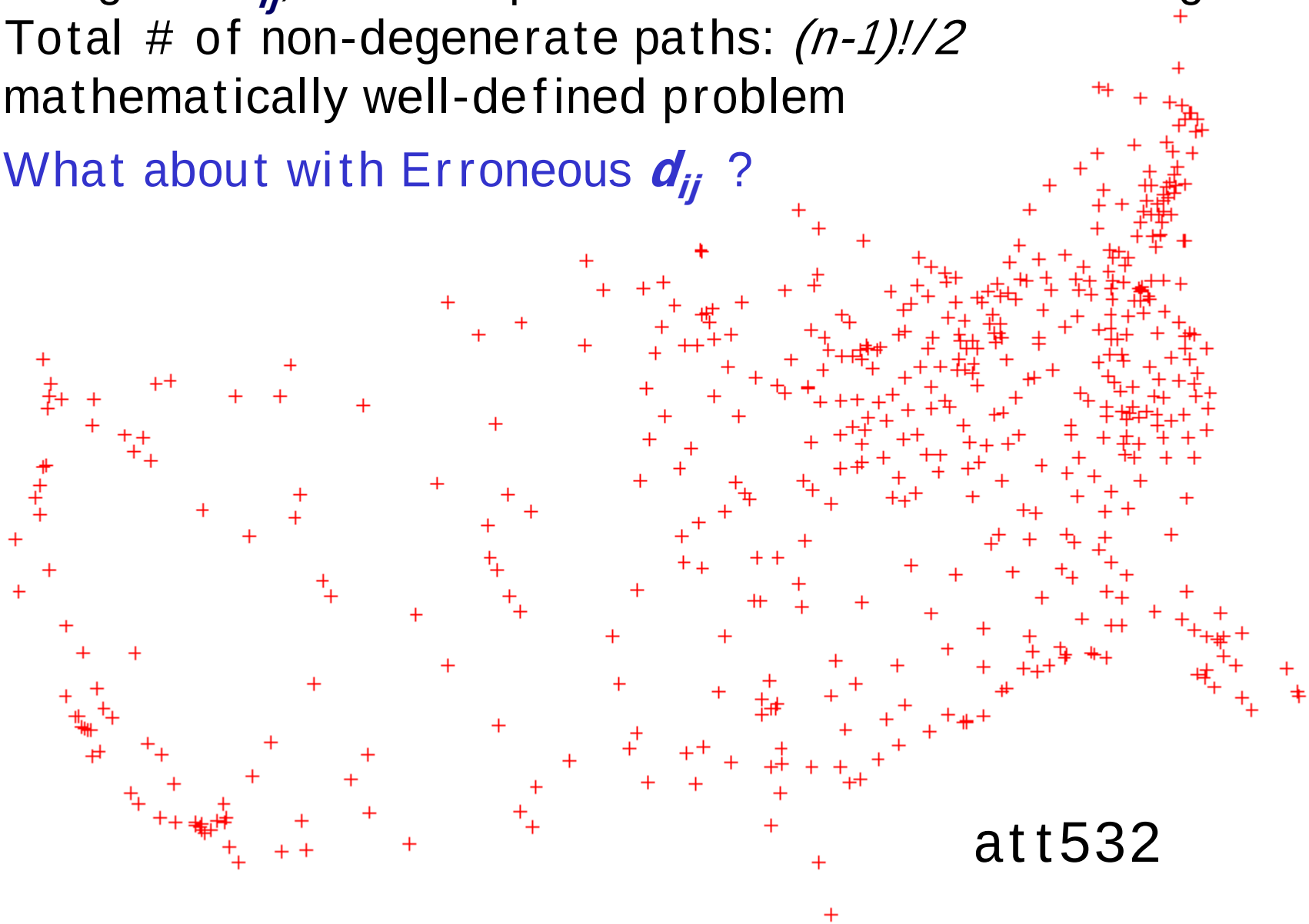
.

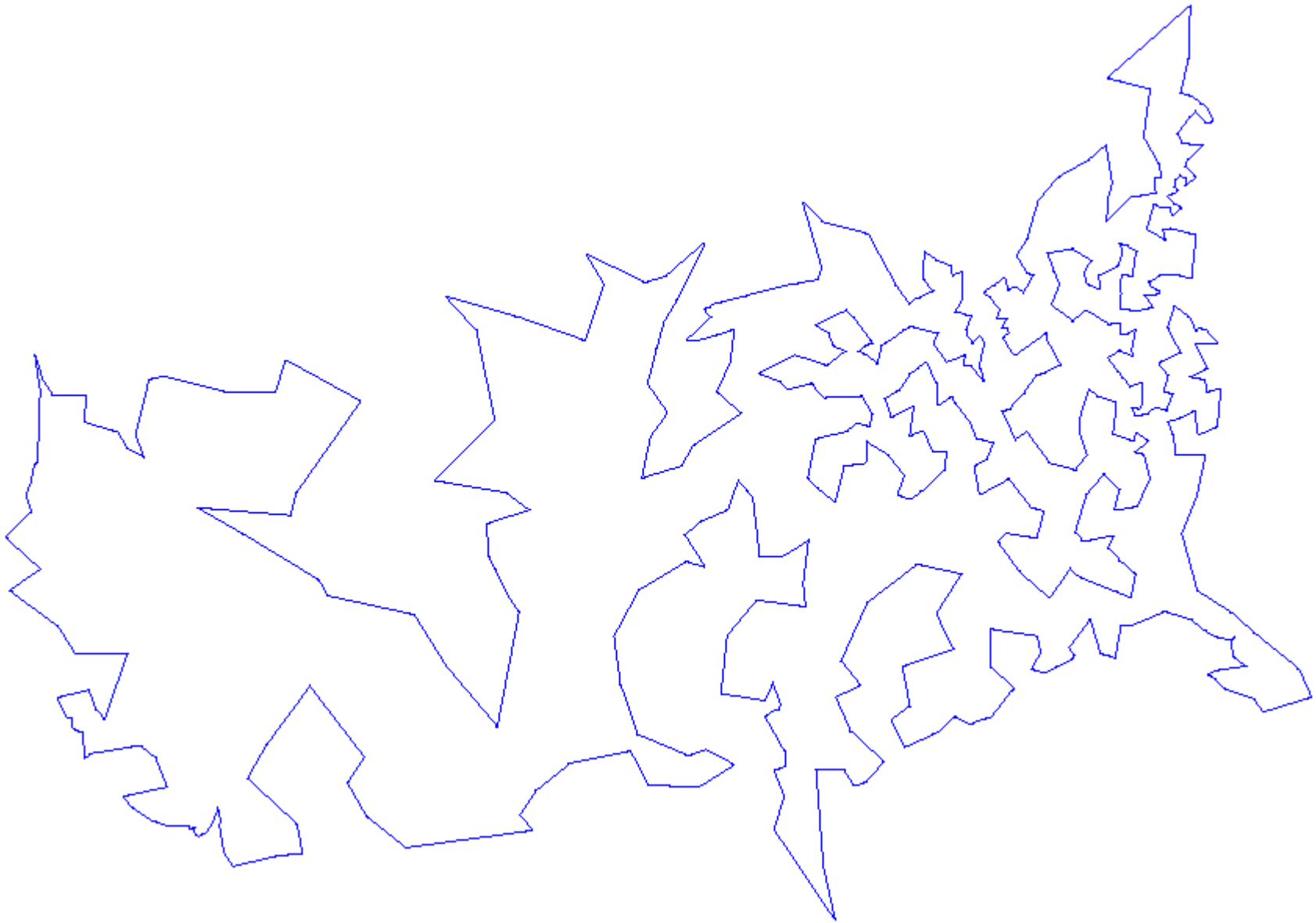
Protein Folding Problem

1. 단백질(protein) 왜 중요한가?
2. 단백질 합성(protein synthesis)과 Central Dogma
3. **Protein folding problem vs. Traveling salesman problem**

Traveling Salesman Problem

- For given d_{ij}, find the path of the shortest tour length
- Total # of non-degenerate paths: *(n-1)!/2*
- mathematically well-defined problem
- What about with Erroneous d_{ij} ?





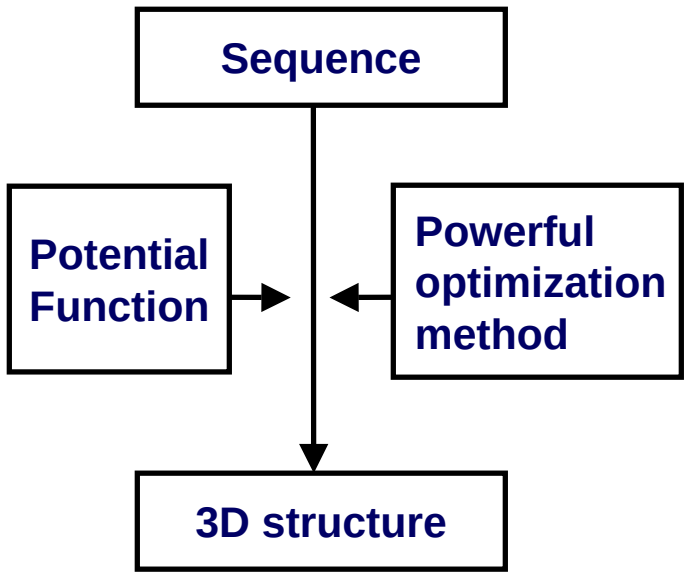
E(x)=...

E(x)=...

E(x)=...

E(x)=...

Ab initio protein folding by computer simulation



Protein Folding Problem

1. 단백질(protein) 왜 중요한가?
2. 단백질 합성(protein synthesis)과 Central Dogma
3. Protein folding problem vs. Traveling salesman problem
4. **Ab. initio protein structure prediction**

CASP1
CASP2
CASP3
CASP4 ✔
Local services
Other links
People

Website index

Hide menu

Protein Structure Prediction Center

Biology and Biotechnology Research Program
Lawrence Livermore National Laboratory, Livermore, California, USA

Welcome to the Protein Structure Prediction Center!

Our goal is to help advance the methods of identifying protein structure from sequence. The Center has been organized to provide the means of objective testing of these methods via the process of blind prediction. In addition to support of the CASP meetings our goal is to promote an objective evaluation of prediction methods on a continuing basis.

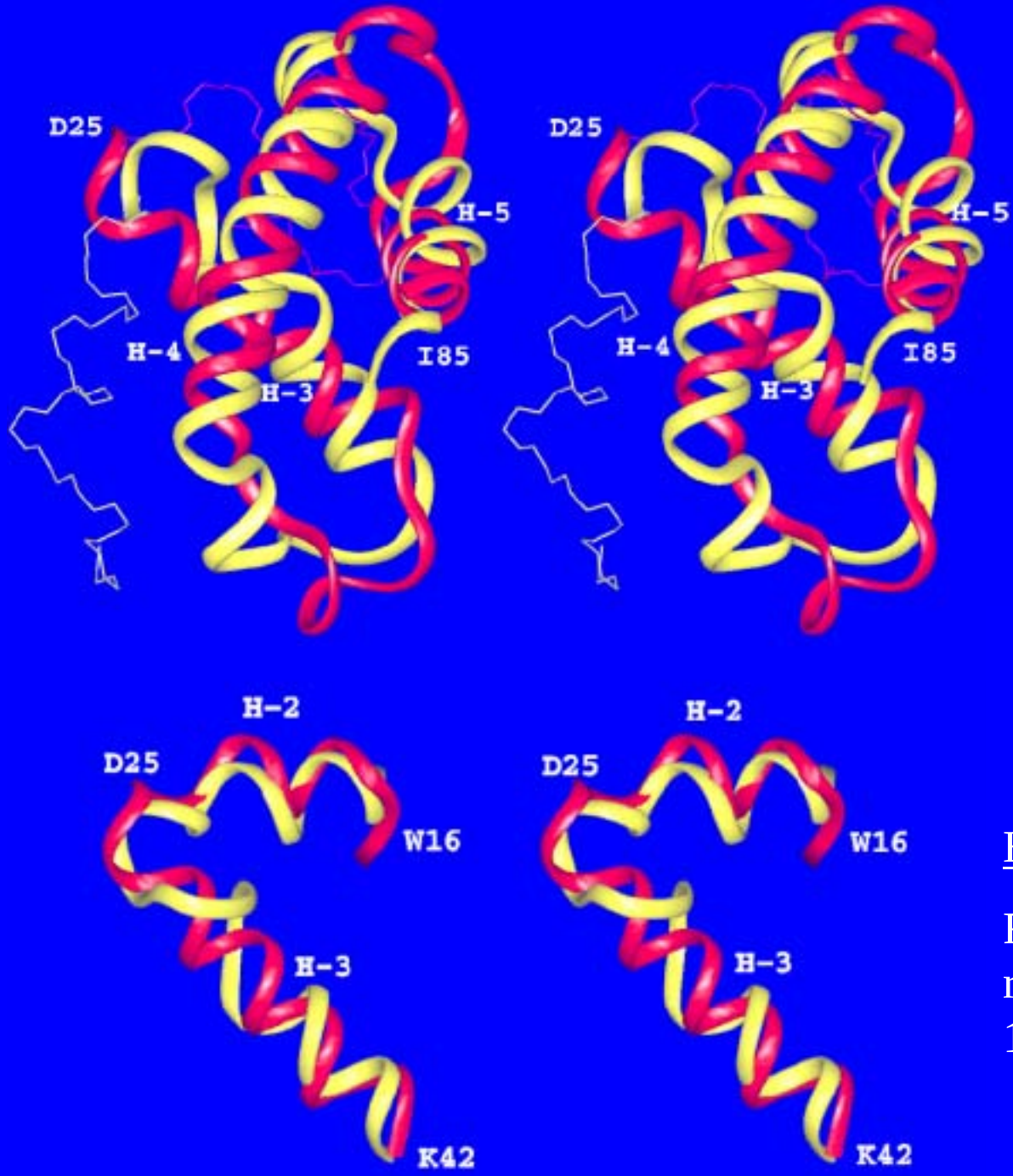
CASP experiment: CASP1 | CASP2 | CASP3 | CASP4

Ten Most Wanted: TMW

The Center, supported by the U.S. Department of Energy, Office of Biological and Environmental Research, is a part of the Biology and Biotechnology Research Program at the Lawrence Livermore National Laboratory.

Local services | Other links | People | Website index

CASP4 protein structure prediction evaluation data



HDEA

RMSD=4.2 Å for 61 residues (80%, residues 25-85)

HDEA Segment

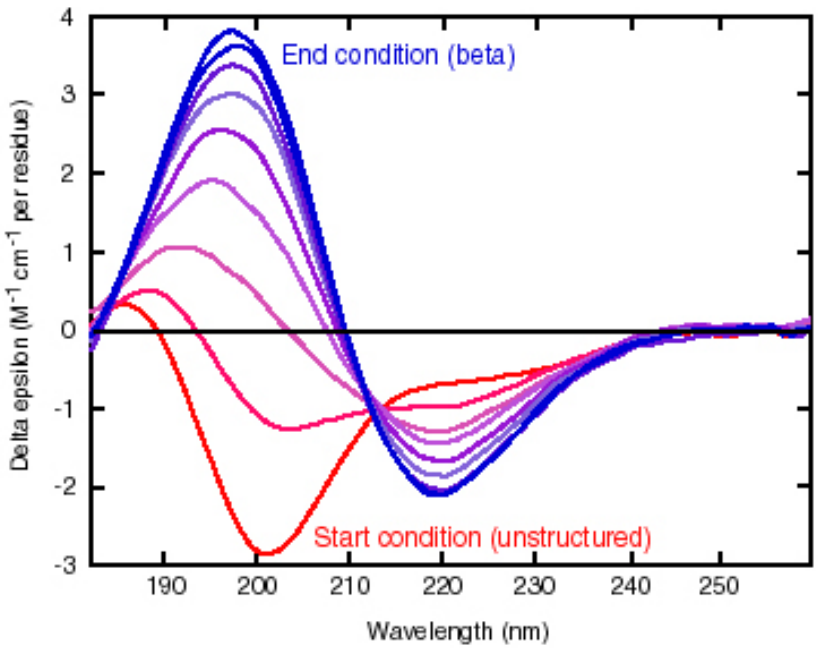
RMSD=2.9 Å for 27 residues (36%, residues 16-42)

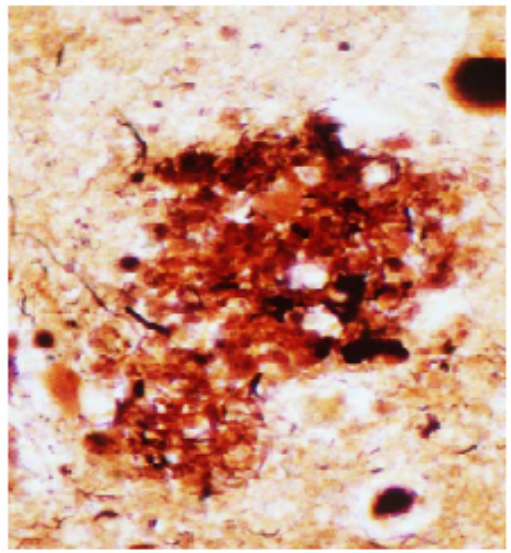
PNAS **96**,5482

Amyloid diseases

A number of diseases (e.g. Alzheimer's, CJD, BSE) involve the folding of proteins and peptides into beta-sheet structures which can polymerise, forming insoluble plaques in nerve tissue (below right).

A model for the Alzheimer's peptide is LRRN, which forms spontaneously into gels with a β-sheet structure.





SRCD spectra* (left) taken during the polymerisation of LRRN peptide show that the rate of polymerisation varies with substitution of a single amino acid residue.

*Collaboration with N.Gay and M. Symmons, Cambridge University

The SRCD data provide important information about the processes involved in polymerisation, and may lead to the development of drugs to treat these diseases.

Protein folding research topics:

- Protein structure prediction
- Protein folding mechanism: MD,MC simulations
- Docking problems
- Secondary structure prediction
- Contact prediction
- Order/disorder prediction
- Multiple sequence alignment
- Potential parameter optimization
- Global optimization of various systems: TSP, molecular clusters, etc.