Protein Structure Prediction and Global Optimization

Jooyoung Lee
http://lee.kias.re.kr
Center for *in-silico* Protein Science
Korea Institute for Advanced Study

2015 Summer School on Polymers in Biology
KIAS, Seoul
June 3 – July 3, 2015

- Contents:
  - How to obtain the ground state
  - Global optimization methods
  - Proteins: What we know and what we don't know
    - Structure prediction/determination
    - Folding mechanism
  - Levinthal's paradox

Examples of Phase Transitions in Nature

- Water-vapor transition (boiling of water)
- Ice-water transition (crystal melting)
- Protein folding

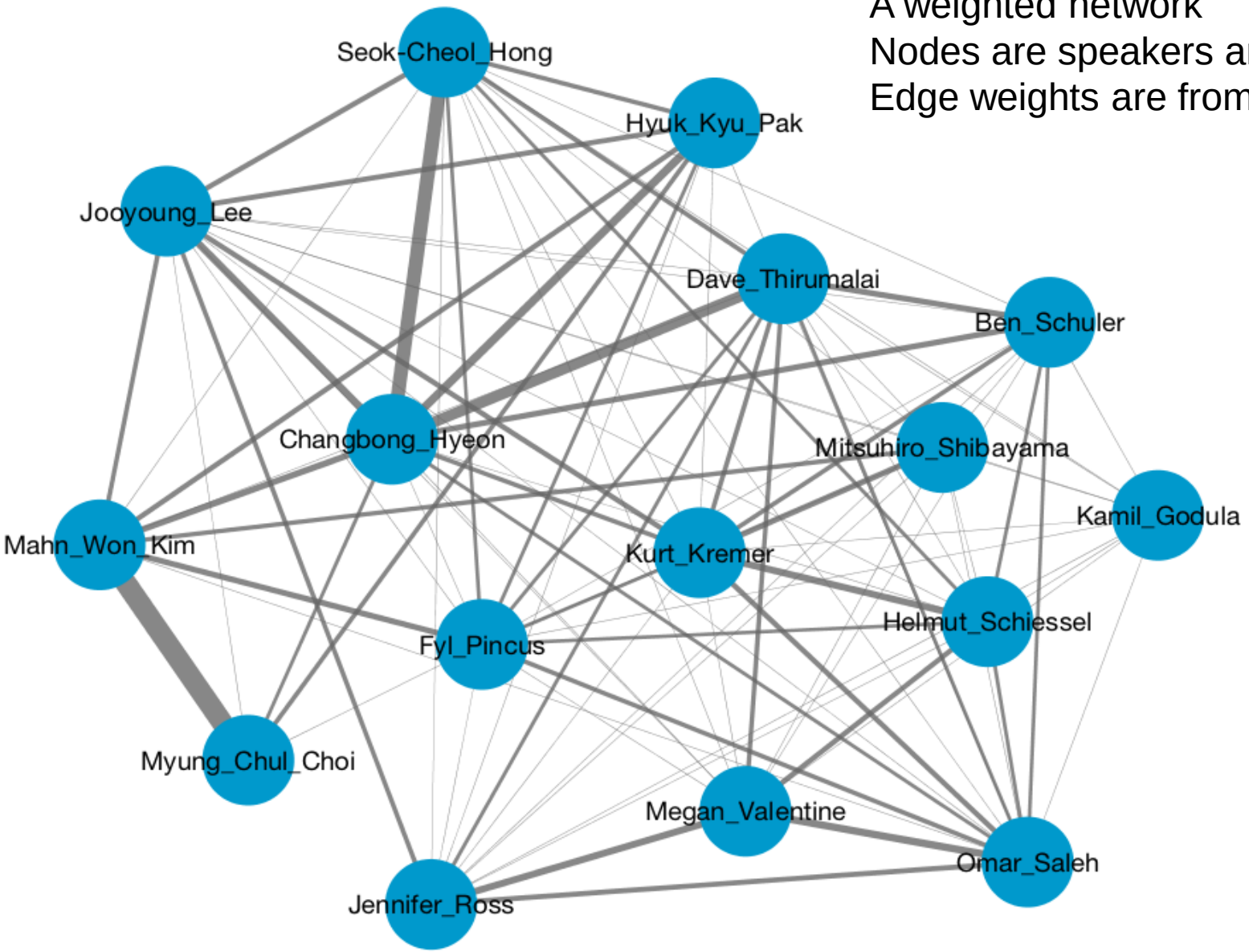
- Spin systems:
  - 2D Ising model
  - Sherrington-Kirkpatrick spin glass

Global Optimization

- Many problems in science and engineering are optimization problems.
- Efficient acquisition of the ground state and/or low-lying excitations is often sufficient to understand the essence of the problem (especially when T<Tc).
- "Prediction" followed by "experimental validation" is one area where theories and experiments can work together.
- Generating 3D atomic models consistent with experimental data is also an important area where computation can contribute significantly ➔ X-ray and NMR protein structure **"determination".**

Speaker's Network of This School

A weighted network
Nodes are speakers and organizers
Edge weights are from Google



Three topics to cover

(1) Community detection of a network by modularity optimization

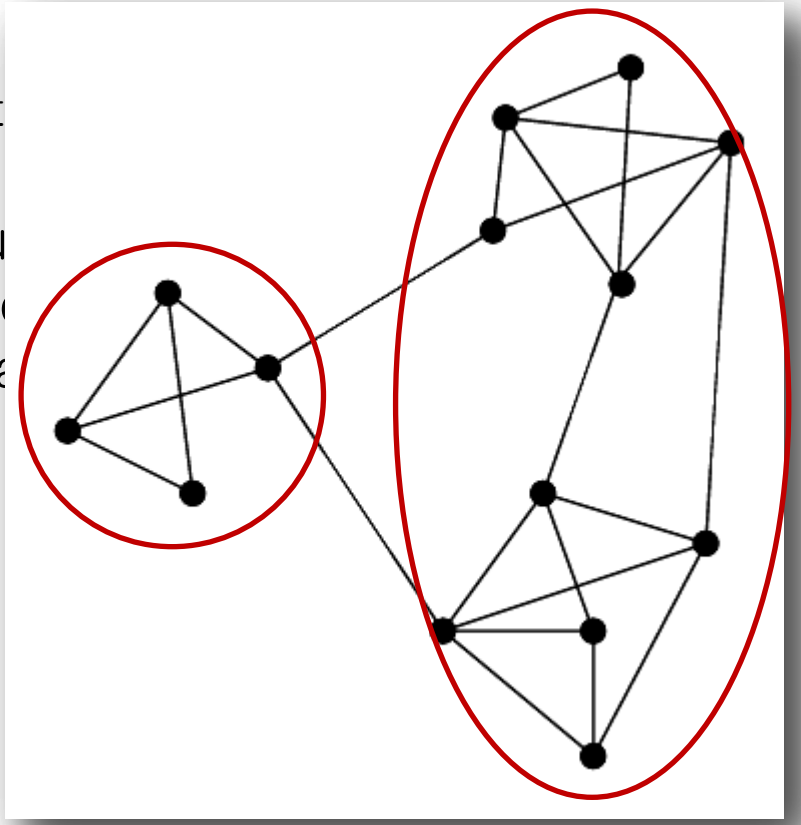
(2) Materials design: Direct bandgap silicon crystal

(3) Protein structure prediction and NMR protein structure determination:
- Using NOE and DHI restraints data from experiments
- Protein structure modeling using sparse & ambiguous NOE restraints

"Community/Module Detection" by Modularity Optimization

- Divide a network into sub-graphs/modules
  - nodes are more densely connected internally
- The most commonly used objective function to evaluate the quality of partition is Q proposed by Girvan and Newman

$$Q = \sum_{s=1}^{r} \left[\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right]$$



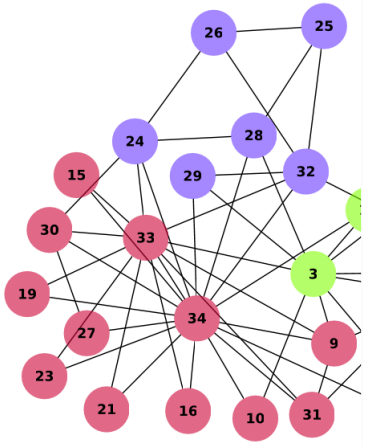
l_s : Number of intra-community edges in s

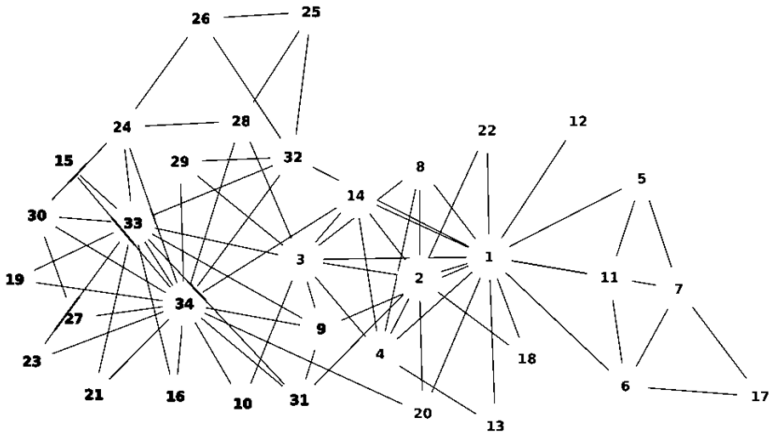
d_s : Sum of degrees of nodes in s

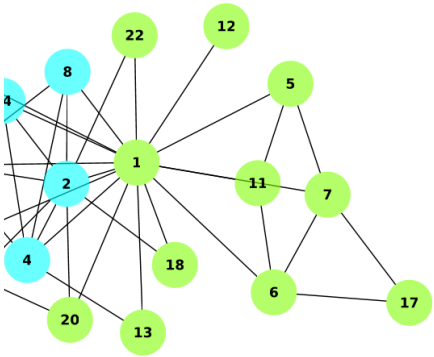
L : Total number of edges in a network

Which one is better?

Zachary's karate club network
Friendship network of members







Q=0.420 Q=0.379

We need an objective function!

Two Issues with Modularity Q

(1) Difficulty of the problem:

- Finding the best Q partition is a hard combinatorial optimization problem (NP-hard)
- The current best stochastic optimization method is simulated annealing (SA)

(2) Relevance of the objective function:

- Is a higher-Q solution more useful to extract hidden information from a network?
- "So far, most works in the literature on graph clustering focused on the development of new algorithms, and **applications were limited** to those few benchmark graphs that one typically uses for testing" from *Community Detection in Graph*s (2010), Physics Report

Benchmark Test of Q-Optimization

Benchmark Test #2: real-world networks

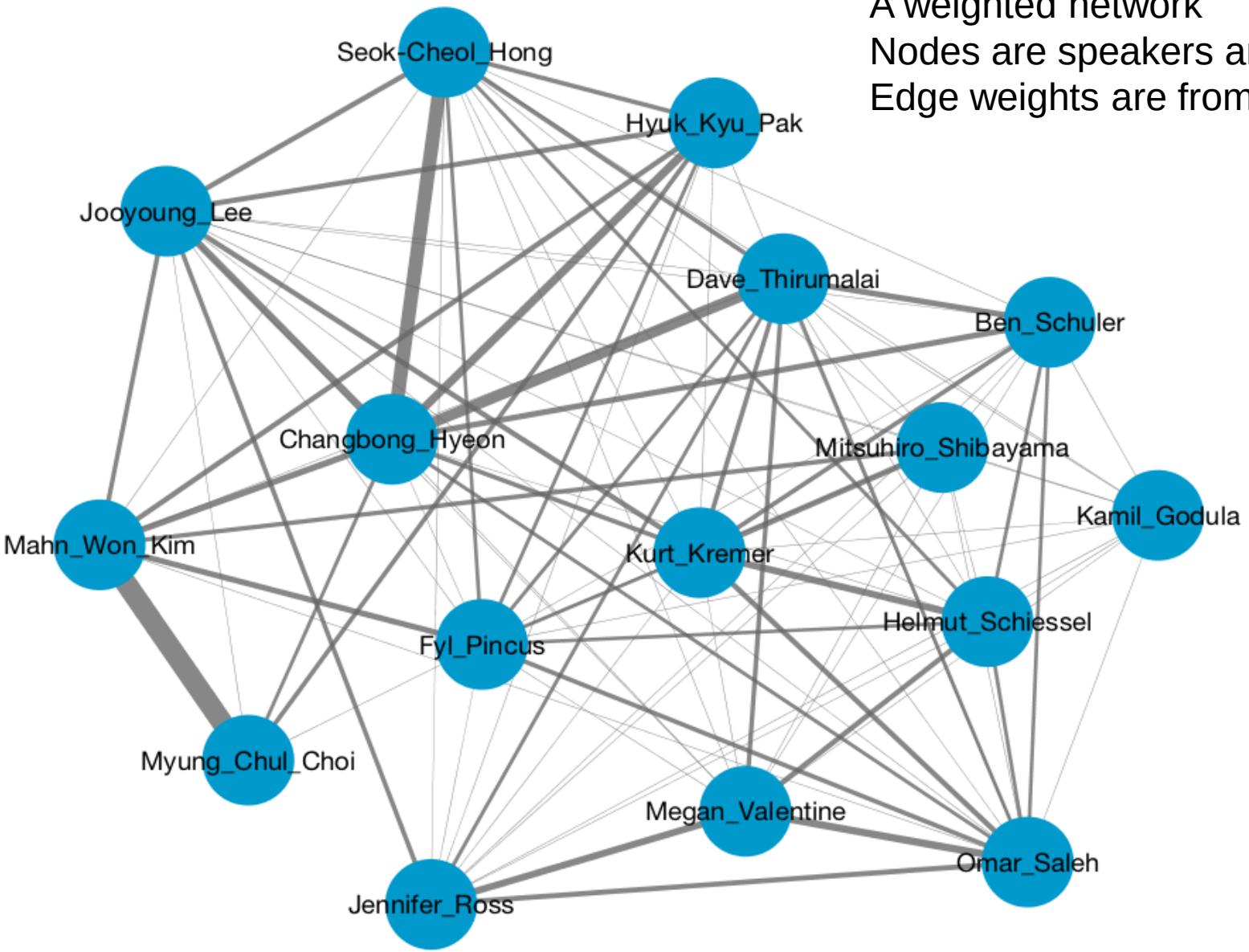
PRE **85, 056702 (2012)**

Nodes	Edges	Network	CSA N_c	CSA Q_{max}	Q_{pub}	Q_{opt}	$\%_{opt}^{SA}$	Source
62	159	Dolphins	5	0.52852	0.5285	0.5285	16.0	[25–27]
77	254	Les Miserables	6	0.56001	0.5600	0.5600	20.0	[27]
105	441	Political books	5	0.52724	0.5272	0.5272	100.0	[26–28]
115	613	College football	10	0.60457	0.6046	0.6046	100.0	[26, 27, 29]
198	2742	Jazz	4	0.44514	0.4451	-	-	[26, 28, 30, 31]
332	2126	USAir97	6	0.36824	0.3682	0.3682	0.0	[27]
379	914	Netscience_main	19	0.84859	0.8486	0.8486	0.0	[27]
453	2025	C. *elegans*	9	0.45325	0.452	-	-	[32]
512	819	Electronic Circuit (s838)	16	0.81936	0.8194	0.8194	0.0	[27]
1133	5451	E-mail	10	0.58283	0.582	-	-	[32]
6927	11850	Erdos02	40	0.71843	0.7162	-	-	[28]
10680	24316	PGP	100	0.88674	0.8841	-	-	[28, 32]
27519	116181	condmat2003	80	0.76745	0.761	-	-	[29]

TABLE III. Comparison between the maximum modularity values obtained by CSA, Q_{max}, with previously published ones, Q_{pub}, and the maximum values obtained by the exact method [27], Q_{opt}, is displayed. N_c denotes the number of communities found by CSA. Source indicates the reference that the modularity value is collected. $\%_{opt}^{SA}$ denotes the percentage of SA runs that reached to the optimal modularity community structure.

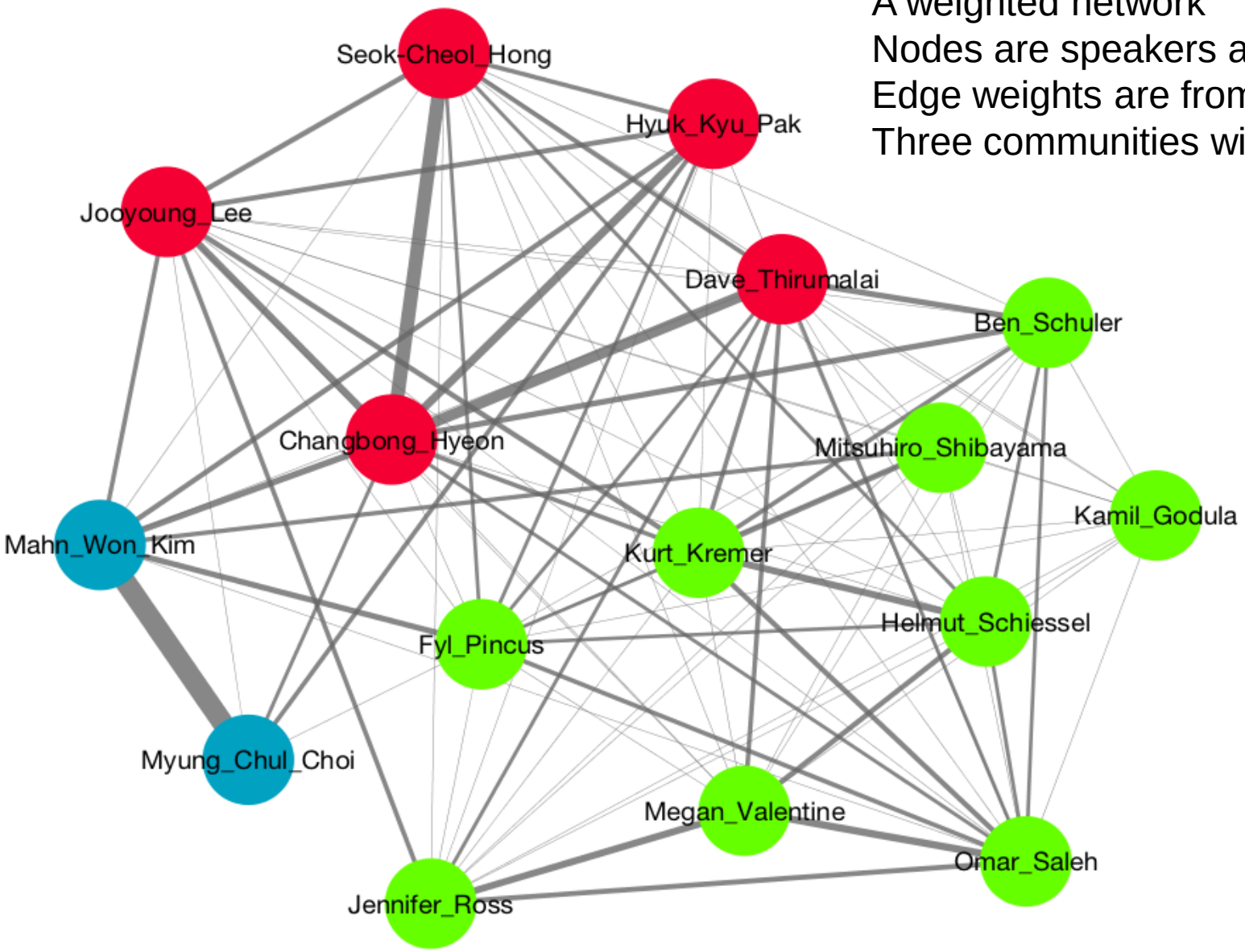
Speaker's Network of This School

A weighted network
Nodes are speakers and organizers
Edge weights are from Google



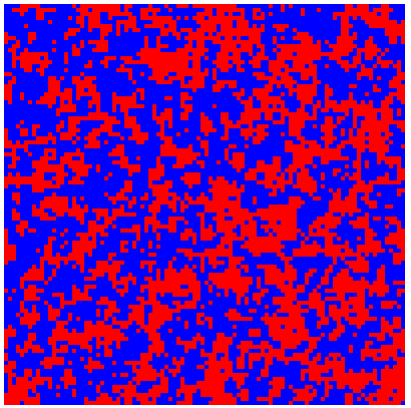
Speaker's Network of This School

A weighted network
Nodes are speakers and organizers
Edge weights are from Google
Three communities with Q = 0.36

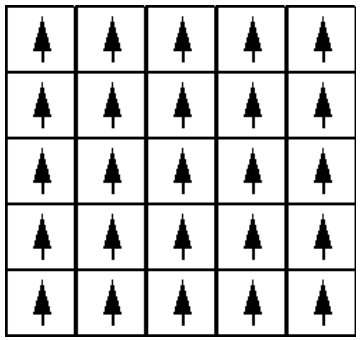


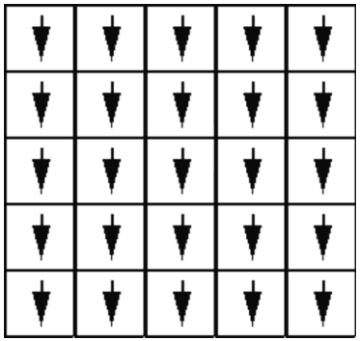
Ferromagnetic Ising model on a 2D square lattice

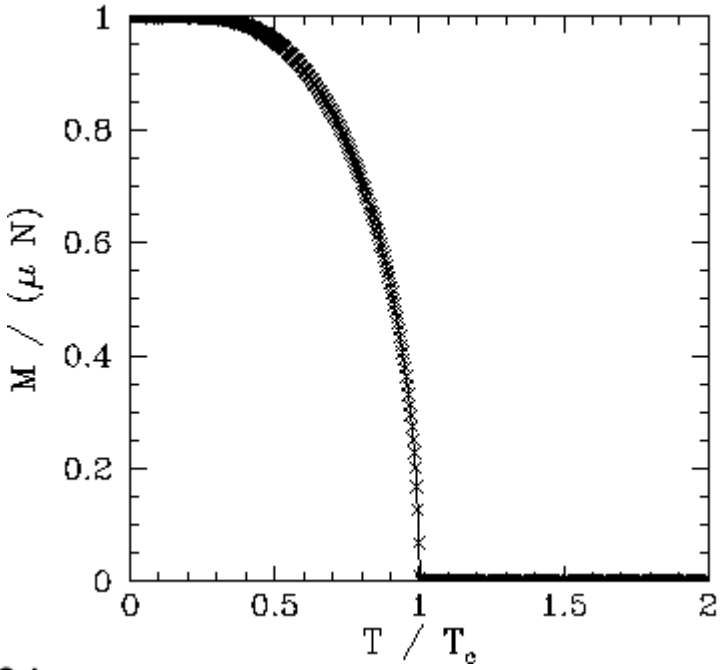
- Spin σ can be up (+1) or down (-1)
- If J > 0, ferromagnetic interaction
- Interaction is only between nearest neighbors
- T < Tc, magnetization is non-zero
- T > Tc m=0



$$\mathcal{H} = -J \sum_{\langle i,j \rangle} \sigma_i \sigma_j \qquad (1)$$



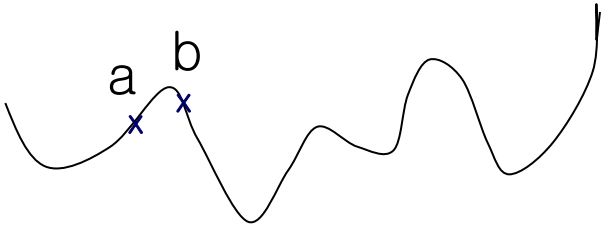




Monte Carlo Simulation of a Canonical System

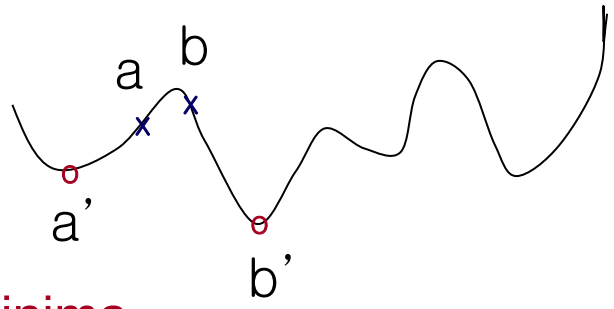
- Each microstate x contributes to the probability of the system with the Boltzmann factor of $P(x) \propto \exp[-\beta E(x)]$ where β is $1/kT$
- Conditions to simulate a canonical system:
  - system should be ergodic
  - detailed balance: transition probability, $T(i \rightarrow j)$, between microstates i and j obey $Pi\ T(i \rightarrow j) = Pj\ T(j \rightarrow i)$
  - Metropolis algorithm:
    - transition to a lower energy state is always accepted
    - transition to a higher energy state is accepted with the probability of $\exp(-\Delta E/kT)$

Reviews on some optimization methods



I. Simulated annealing (SA)

- Mimics the natural process of crystal forming from magma
- Generates **MC moves** as T is gradually lowered from high T to low T
- The most general method



II. MC with minimization (MCM)

- Considers only local minima
- **MC moves** in the solution space of local minima
- Quite successful method!
- PNAS Vol. 84, 6611-6615 (1987); J. Phys. Chem. A 101, 5111-5116 (1997) (Basin-Hopping Method).

III. Genetic algorithm (GA)

- Pool of conformations (generation)
- Subsequent generations are obtained by mating and mutation of parent generation by natural selection of fitness function (evolution of species)

P1. Sherrington-Kirkpatrick spin glass

- Consider N Ising spins interacting with all the other spins.
- Interaction coupling J is randomly assigned as either +1 or -1.
- Hamiltonian is $\mathcal{H} = -\Sigma_{\langle ij \rangle} \frac{J_{ij}}{\sqrt{N}} S_i S_j$

- The goal is to find the ground state for a given realization of randomness.
- Try with N=511 and generate a realization of SK spin glass. Find the lowest energy configuration using a global optimization method (SA or MCM)
- For 100 realizations of N=511 SK spin glass, obtain the average value and the standard deviation of the ground state energy. Is the average value close to - 0.75238?
- Reference: Ground-state energy and energy landscape of the Sherrington-Kirkpatrick spin glass, Phys.Rev.B, Vol. 76, 184412 (2007).

Simulated Annealing for the SK spin glass

1. N=511;T=100;beta=1/T;iter=0;Emin=9999
2. Generate a random realization of Jij between all pairs {i,j} using a random number generator

   DO i=1,N-1

   DO j=i+1,N

   if ran() > 0.5 then Jij=1, else Jij=-1

   Jji=Jij

   ENDDO

   ENDDO
3. Generate a random configuration of Spin(i)=+1 or -1 as above
4. Calculate the total energy→ E;if E ≤Emin then (Emin=E and save the current Spin())
5. Perturb the current configuration Spin()) by the single random spin flip move; iter=iter+1;if iter > 100 then (beta=beta/0.99 and iter=0)
6. Calculate the total energy of the new configuration → Enew
7. Calculate p=exp {(E-Enew)*beta}
8. if ran() > p then GOTO 5
9. accept the move and set E=Enew
10. if E ≤Emin then (Emin=E and save the current Spin())
11. GOTO 5

MCM for SK the spin glass

1. N=511;T=10;beta=1/T;Emin=9999;iter=0;iterm=1000
2. Generate a random realization of Jij
3. Generate a random configuration of Spin(i)=+1 or -1, and calculate E
4. Perform quenching N times using the single spin flip move
5. If E is changed GOTO 4
6. Quenching is finished
7. if E ≤Emin then (Emin=E and save the current Spin())
8. If iter>iterm then goto 17;Perturb the current configuration Spin()) by “multiple” random spin flip move
9. Perform quenching N times using the single spin flip move
10. If E is changed GOTO 9
11. Quenching is finished; iter=iter+1
12. Calculate p=exp {(E-Enew)*beta}
13. if ran() > p then GOTO 8
14. accept the move and set E=Enew
15. if E ≤Emin then (Emin=E and save the current Spin())
16. GOTO 8
17. END

Three topics to cover

(1) Community detection of a network by modularity optimization

(2) Materials design: Direct bandgap silicon crystal

(3) Protein structure prediction and NMR protein structure determination:
- Using NOE and DHI restraints data from experiments
- Protein structure modeling using sparse & ambiguous NOE restraints

Issues with direct gap silicon crystal

- Silicon based materials are cheep → most solar cell materials are all silicon based.
- Existing silicon crystals are either metallic or of indirect band gap in their electronic structures.
- Indirect band gap material are inefficient as a solar-cell material.
- Goal is to design crystalline silicon with direct band gap.

Global Optimization by CSA

- We optimized the internal energies of our crystalline Si structures by using first-principles quantum calculations, calculated their electronic structures *ab initio,* and used these structures to select direct-bandgap solutions using CSA
- For a given number of silicon atoms in a unit cell, we optimize the band gap property using the unit cell shape and size and the 3D positions of silicon atoms as variables

Direct band gap silicon crystal

TABLE I. For each structure, the lattice type, the number of atoms per unit cell, the volume per atom, the energy per atom relative to diamond Si, the direct gap size (E_g^d), and the indirect gap size (E_g^i) are shown, based on the PBE calculations. Lattice types are abbreviated, such as tc: triclinic, bcm: base-centered monoclinic, or: orthorhombic, pm: primitive monoclinic, bct: body-centered tetragonal, sc: simple cubic, bcc: body-centered cubic, rho: rhombohedral, and fcc: face-centered cubic. Q135 is classified as a quasidirect gap semiconductor according to the quasiparticle calculation, while it is of direct gap according to the PBE functional. All eight direct gap structures shown in the top eight rows are confirmed as direct gap semiconductors in both calculations.

Structure	Lattice	Atoms	(Å^3/atom)	(eV/atom)	$\bar{r}$ (Å), σ_r (Å)	$\bar{\theta}$ (°), σ_θ (°)	E_g^d (eV)	E_g^i (eV)	Space group	Ref.
D262	pm	10	21.02	0.08	2.37, 0.04	109.26, 8.17	**0.29**		$P2_1/m$ (*No.* 11)	
D12	or (C)	10	21.56	0.13	2.37, 0.01	108.98, 11.19	**0.50**		$Cmmm$ (*No.* 65)	
D239	tc	10	22.72	0.16	2.37, 0.03	108.69, 13.50	**0.77**		$P1$ (*No.* 1)	
D63	bcm	12	21.10	0.12	2.37, 0.04	109.09, 9.76	**0.66**		$C2/m$ (*No.* 12)	
D135	bcm	12	21.24	0.22	2.38, 0.05	108.42, 14.73	**0.64**		Cc (*No.* 9)	
D243	tc	12	21.88	0.29	2.38, 0.04	107.29, 18.42	**0.61**		$P1$ (*No.* 1)	
D76	bcm	20	21.70	0.13	2.37, 0.03	109.01, 10.59	**0.57**		$C2$ (*No.* 5)	
D979	tc	20	21.17	0.29	2.38, 0.05	108.56, 18.40	**0.60**		$P1$ (*No.* 1)	
Q130	bcm	12	21.86	0.08	2.37, 0.02	108.97, 9.78	0.64	**0.63**	$C2/m$ (*No.* 12)	
Q135	bcm	12	21.50	0.15	2.37, 0.04	108.95, 11.78	**0.93**		$C2/c$ (*No.* 15)	

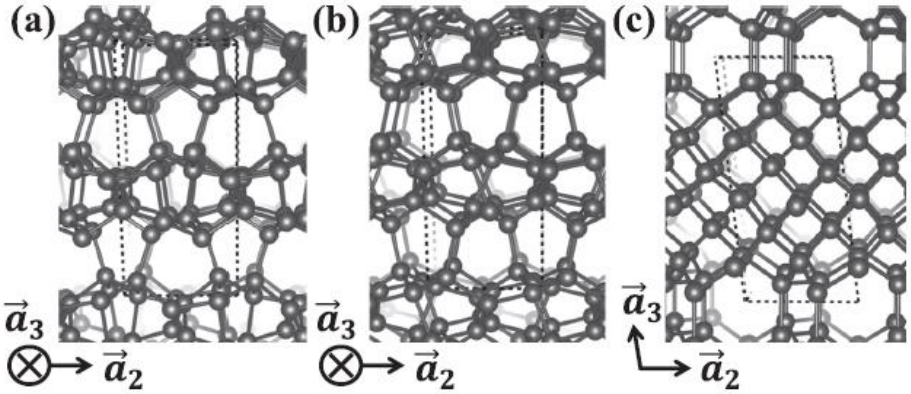


FIG. 3. Atomic structures of (a) D135, (b) Q135, and (c) I926 are shown. No coordination defects are found.

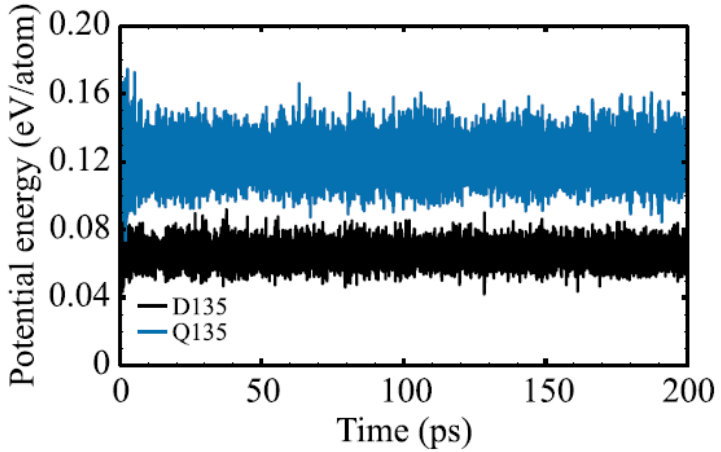


FIG. 6. (Color online) The thermal stability of D135 and Q135 is examined by performing first-principles molecular dynamics simulations for 200 ps at temperatures 500 and 900 K for D135 and Q135, respectively. Potential energy fluctuations are obtained for a supercell containing 96 atoms (eight unit cells).

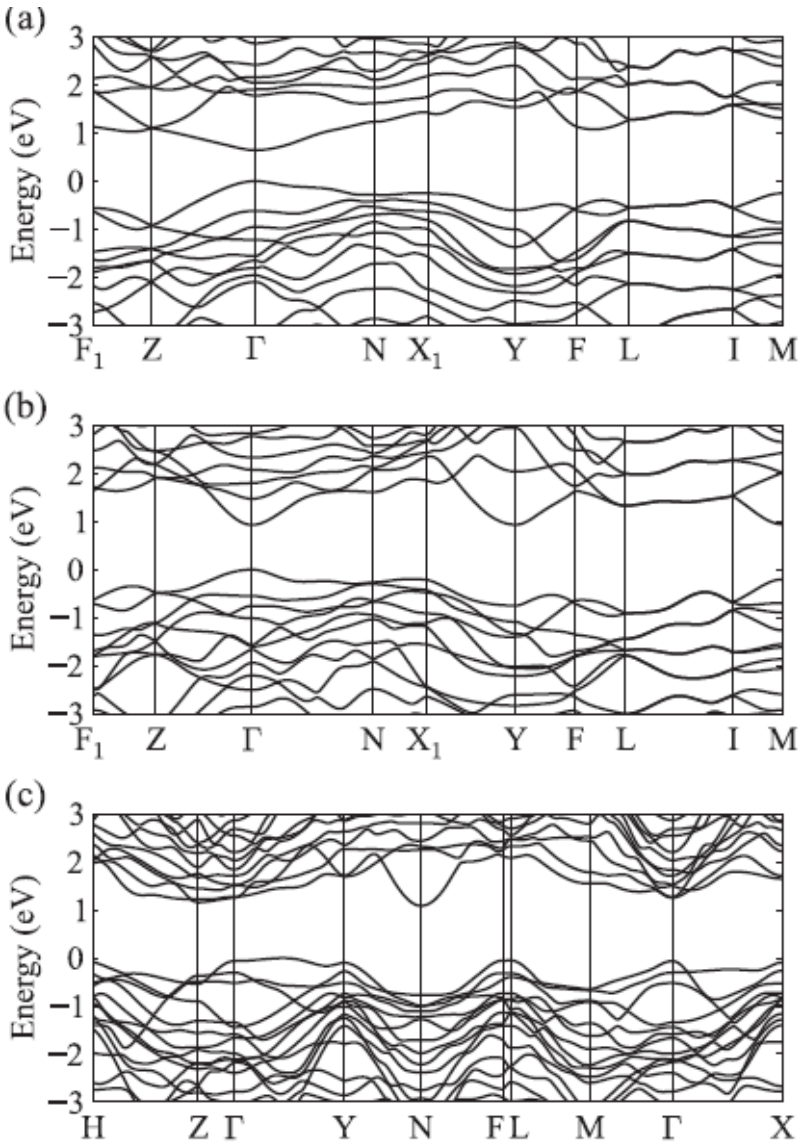


FIG. 7. The PBE electronic band structures of (a) D135, (b) Q135, and (c) I926 are shown near the Fermi level. The Bravais lattices of D135, Q135, and I926 are base-centered monoclinic.

Materials design: direct-band-gap silicon allotropes
Phys. Rev. B 90, 115209 (2014)

to be stable up to 200 ps, which is much longer than previous simulation times of ~10 ps [9].

The band gap variation against external pressure is shown for D135, I926, and diamond silicon in Fig. 4(b). We find that the direct band gap of D135 is fairly stable upon the increase

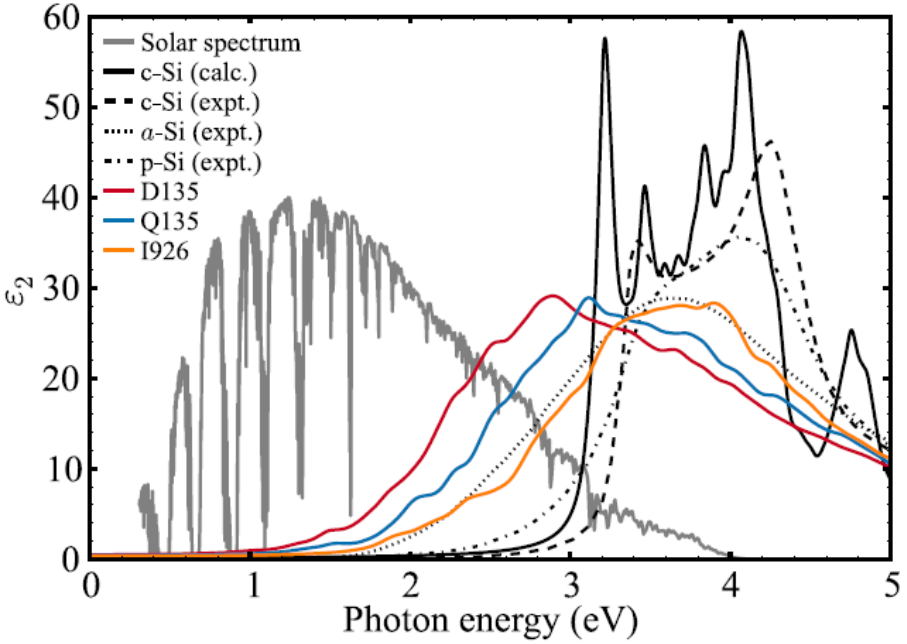


FIG. 8. (Color online) The imaginary part of the dielectric function $\varepsilon_2(\omega)$ is shown as a function of photon energy. Data for D135

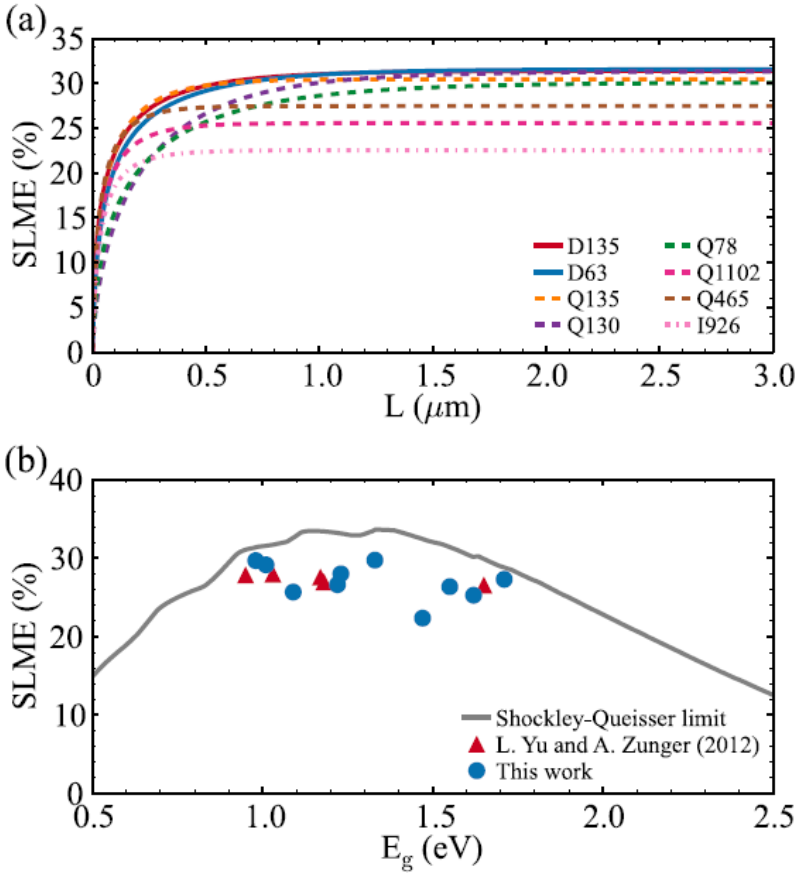


FIG. 9. (Color online) (a) The spectroscopic limited maximum efficiency (SLME) [38] is calculated as the function of film thickness L for D135, D63, Q130, Q135, Q78, Q1102, Q465, and I926. For

Three topics to cover

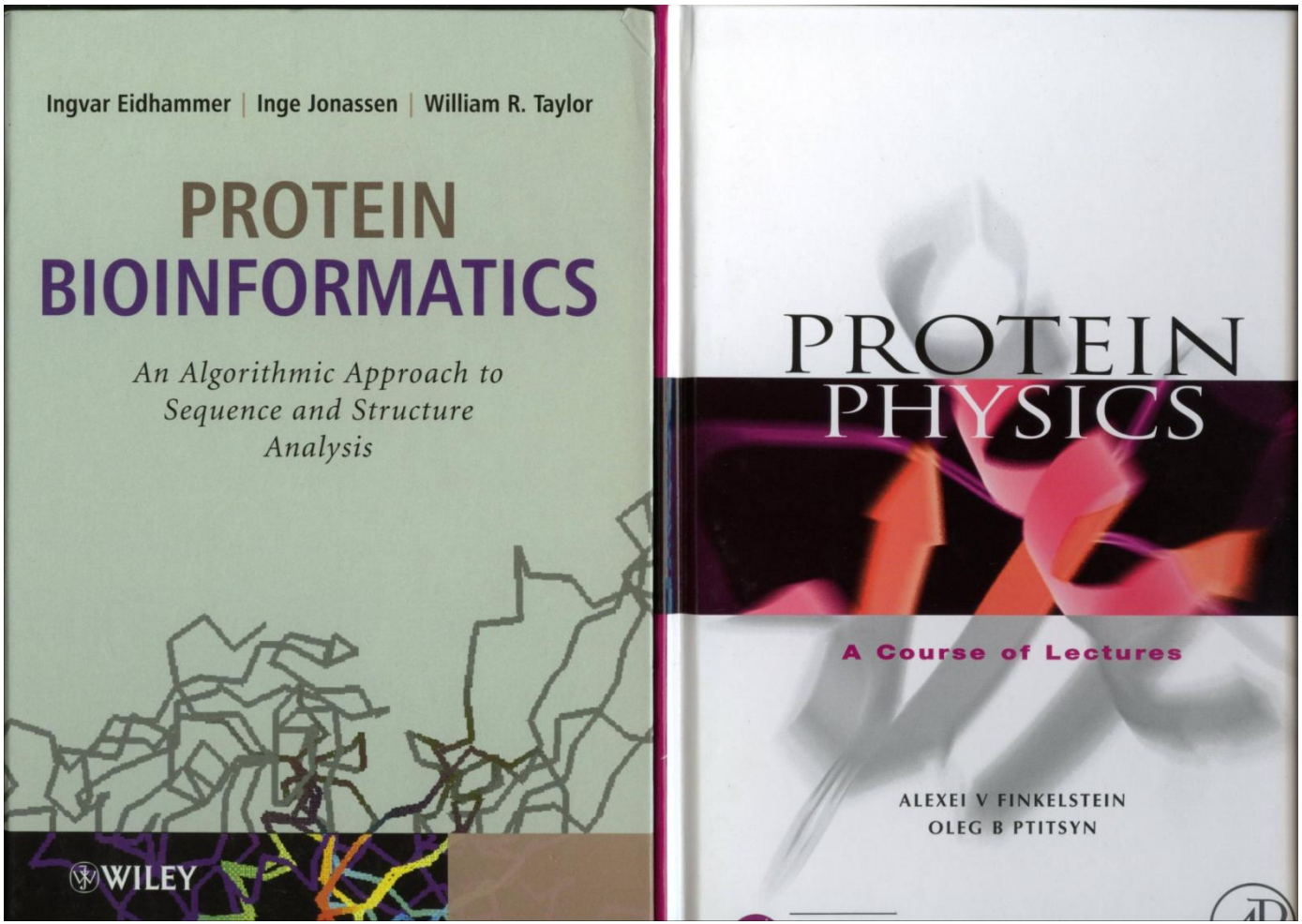
(1) Community detection of a network by modularity optimization

(2) Materials design: Direct bandgap silicon crystal

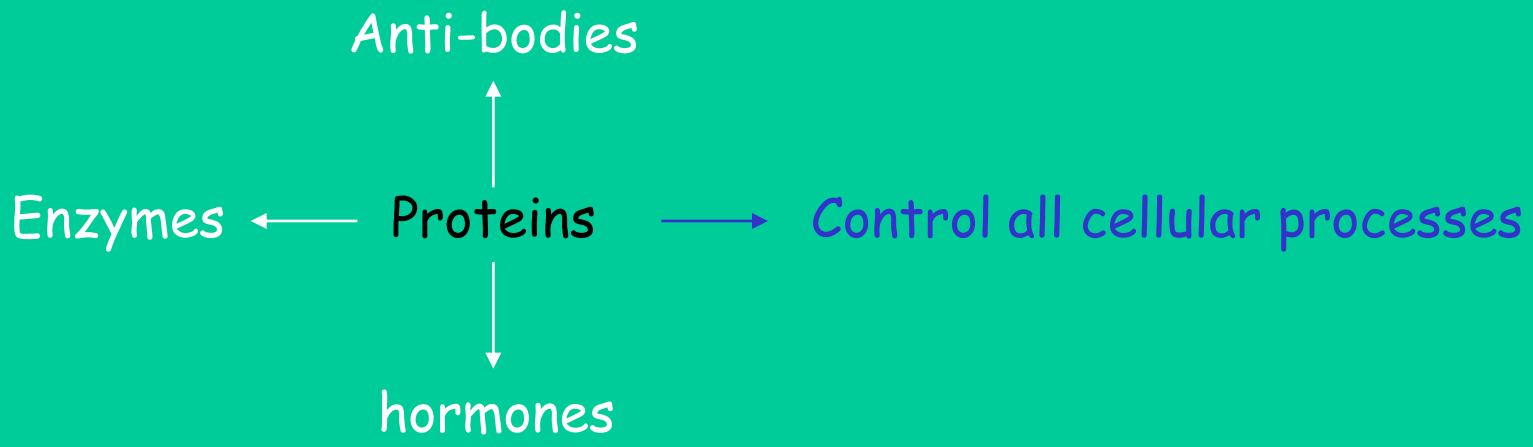
(3) Protein structure prediction and NMR protein structure determination:

- Using NOE and DHI restraints data from experiments
- Protein structure modeling using sparse & ambiguous NOE restraints

2 books about protein structure study for mathematicians/physicists/computer scientists



Proteins are important



sequence (genome) → structure (post-genome) → function

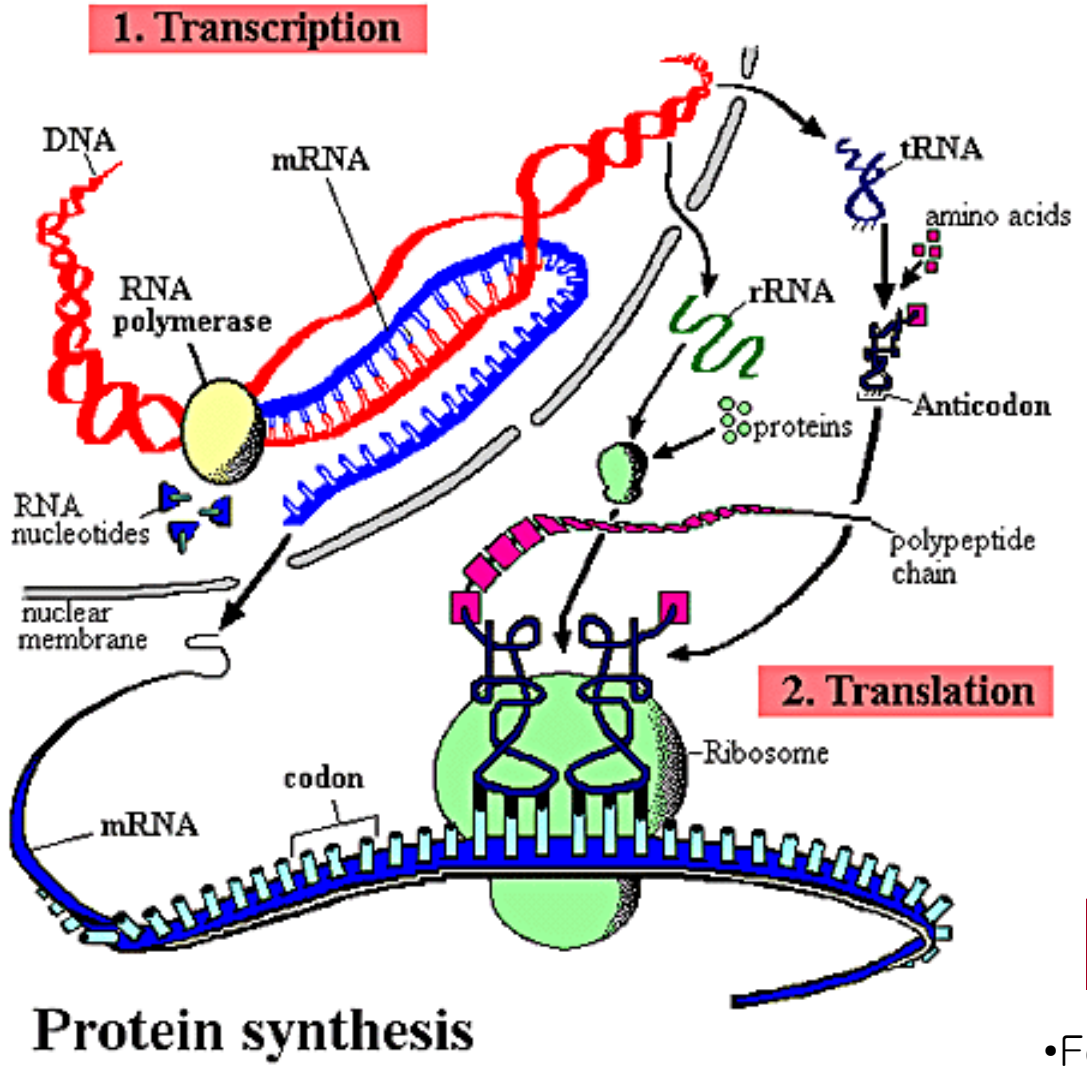
scientific bottleneck

The most challenging problem of this century

(Protein Folding Problem)

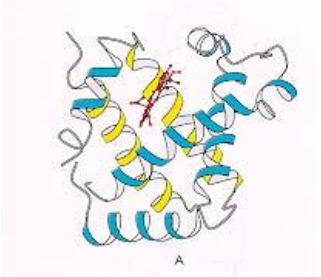
Structure-function relationship of Potassium Channels

- In the field of cell biology, potassium channels are the most widely distributed type of ion channel and are found in virtually all living organisms. They form potassium-selective pores that span cell membranes. Furthermore potassium channels are found in most cell types and control a wide variety of cell functions
- Using X-ray crystallography, profound insights have been gained into ***how potassium ions pass through these channels and why (smaller) sodium ions do not***. The 2003 Nobel Prize for Chemistry was awarded to Rod MacKinnon for his pioneering work in this area.



Protein synthesis

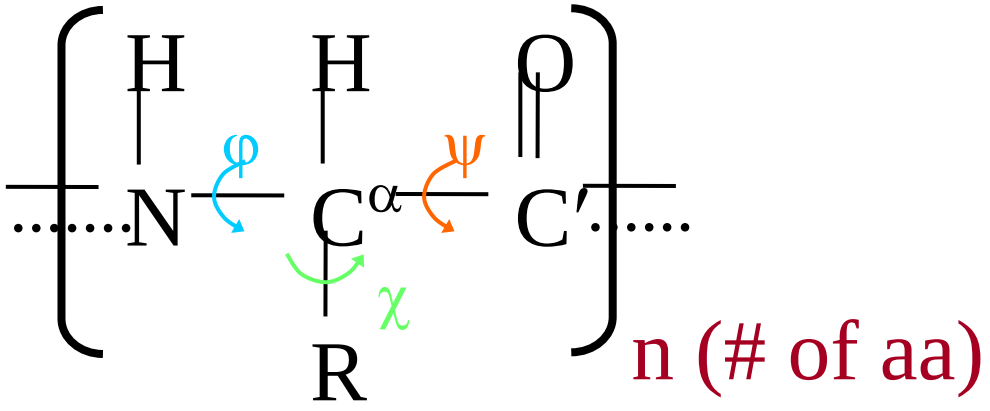
Human hemoglobin



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



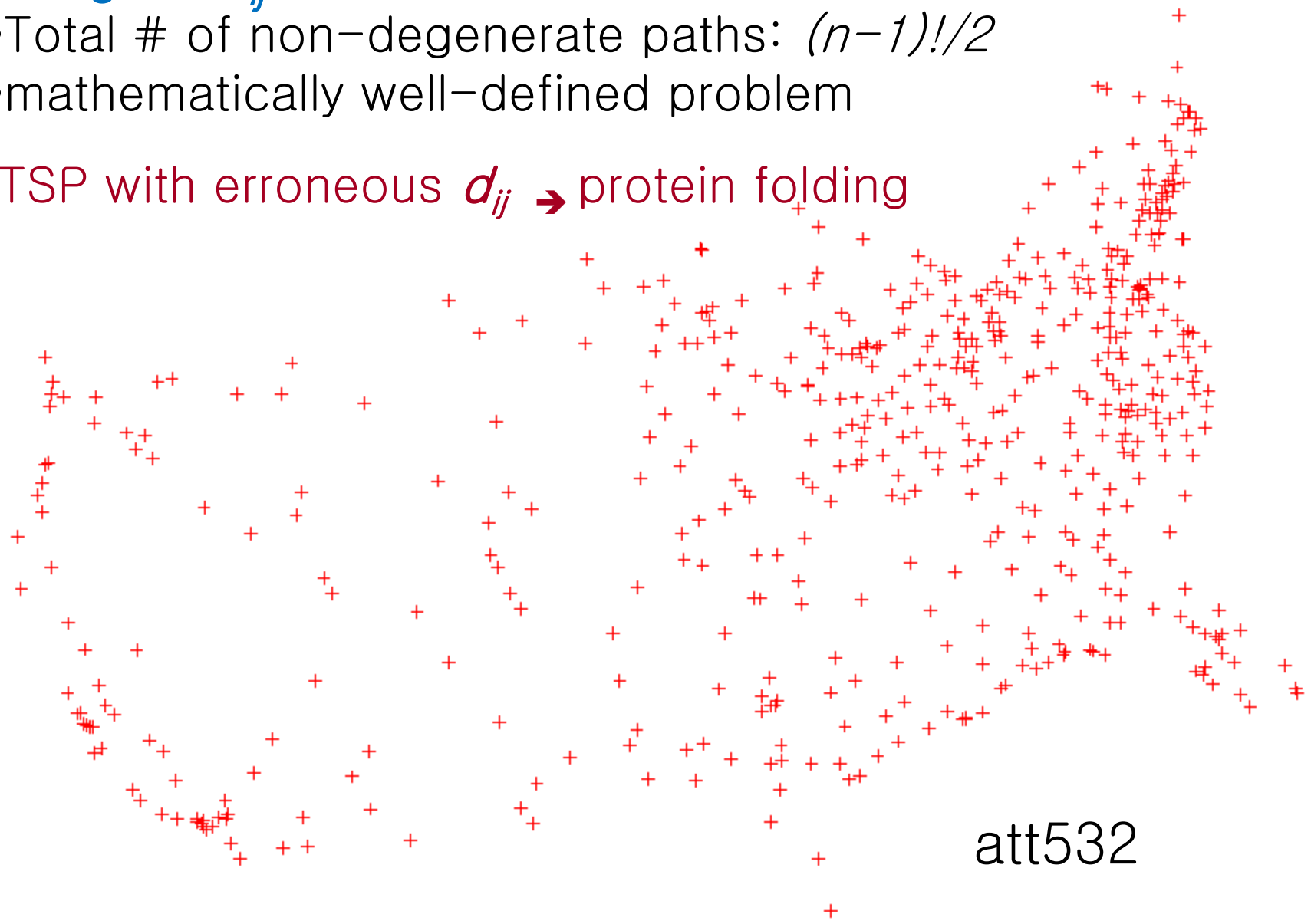
- Proteins are chain molecules made of amino acids.
- There are 20 kinds of aa. R=H,CH3,CH2-Ch2-S-Ch3,…
- # of possible structures of a protein with n aa ~ 10^n
- ~10^5 kinds of proteins in human body.

- Each protein has a unique 3D structure.
- TH by Anfinsen (Science, 181, 223 (1973)): 3D structure of a native protein in its physiological environment is the one in which the free energy of a "whole" system is lowest. (3D structure of a protein is determined by its sequence and its environment)

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

- TSP with erroneous d_{ij} → protein folding



att532

KIAS

e. kr

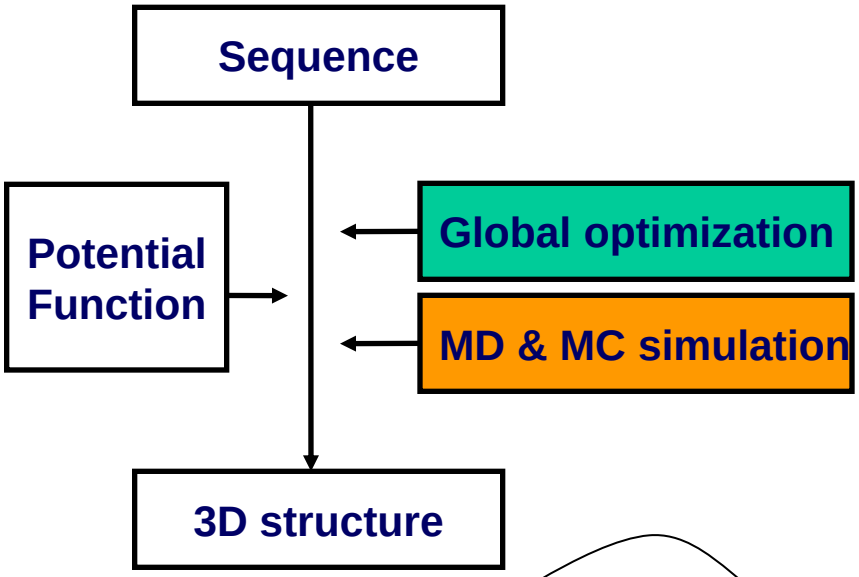
Ab initio protein folding by computer simulation

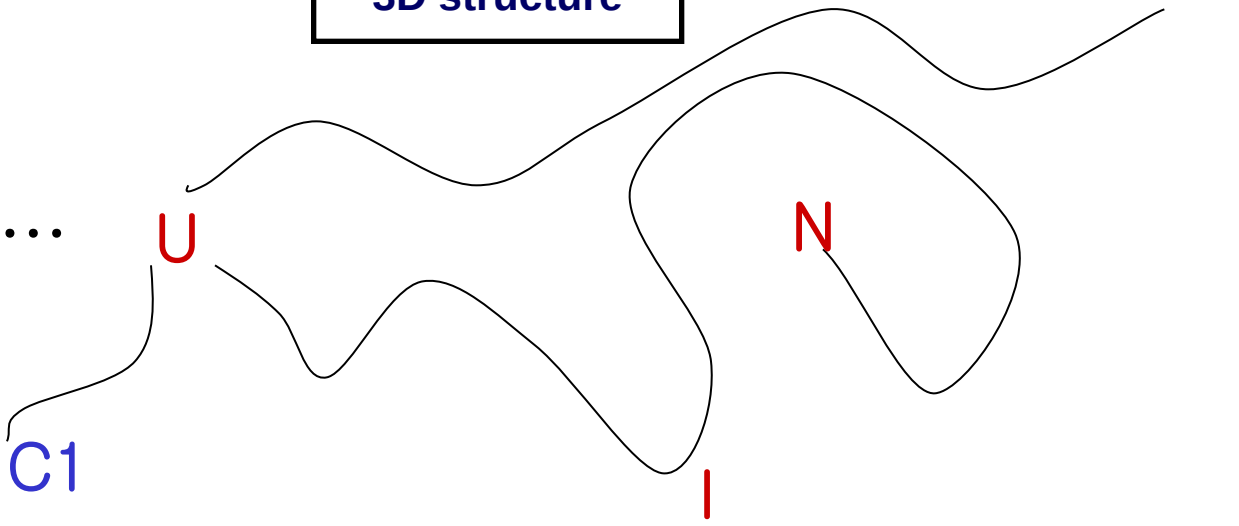
E(x)=…

E(x)=…

E(x)=…

E(x)=…





What we know about proteins:

- Primary structure: The Nobel Prize in Chemistry 1958 was awarded to Frederick Sanger "for his work of 1955 on the structure of proteins (the amino-acid sequence for a protein, insulin)".
- Secondary structure: The Nobel Prize in Chemistry 1954 was awarded to Linus Pauling "for his research into the nature of the chemical bond and its application to the elucidation of the structure of complex substances".
- Tertiary (3D) structure: The Nobel Prize in Chemistry 1962 was awarded jointly to Max Perutz and John Kendrew "for their studies of the structures of globular proteins". Kendrew (1957): myoglobin; Perutz (1959): hemoglobin.
- Quaternary (4D) structure

Two driving forces of protein folding

1. Hydrogen bonding btw NH and CO → alpha-helices and beta sheets.

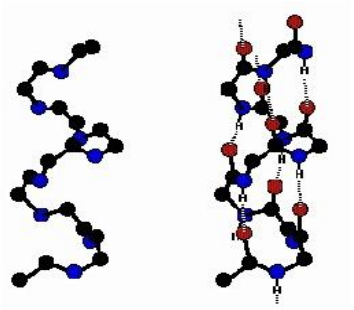
2. Hydrophobic interaction: hydrophobic residues prefer to be inside of a protein and hydrophilic residues on the protein surface.

•Primary Structure

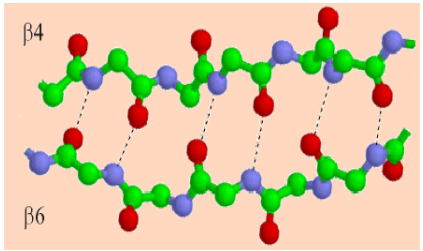
1D sequence of amino acids

e.g.)cys-gly-val-ala-ala

•Secondary Structure



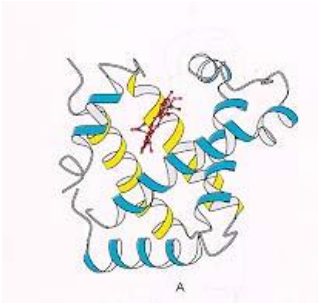
α helix

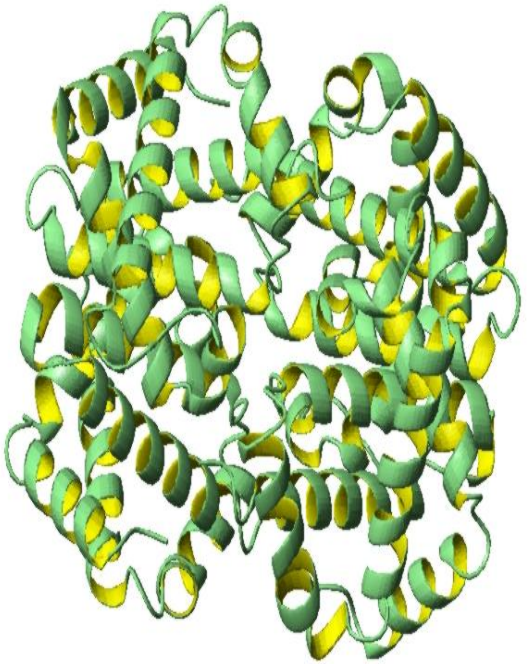


β sheet

•Tertiary Structure

• 3d arrangement





deoxy human hemoglobin

(oxygen transport)

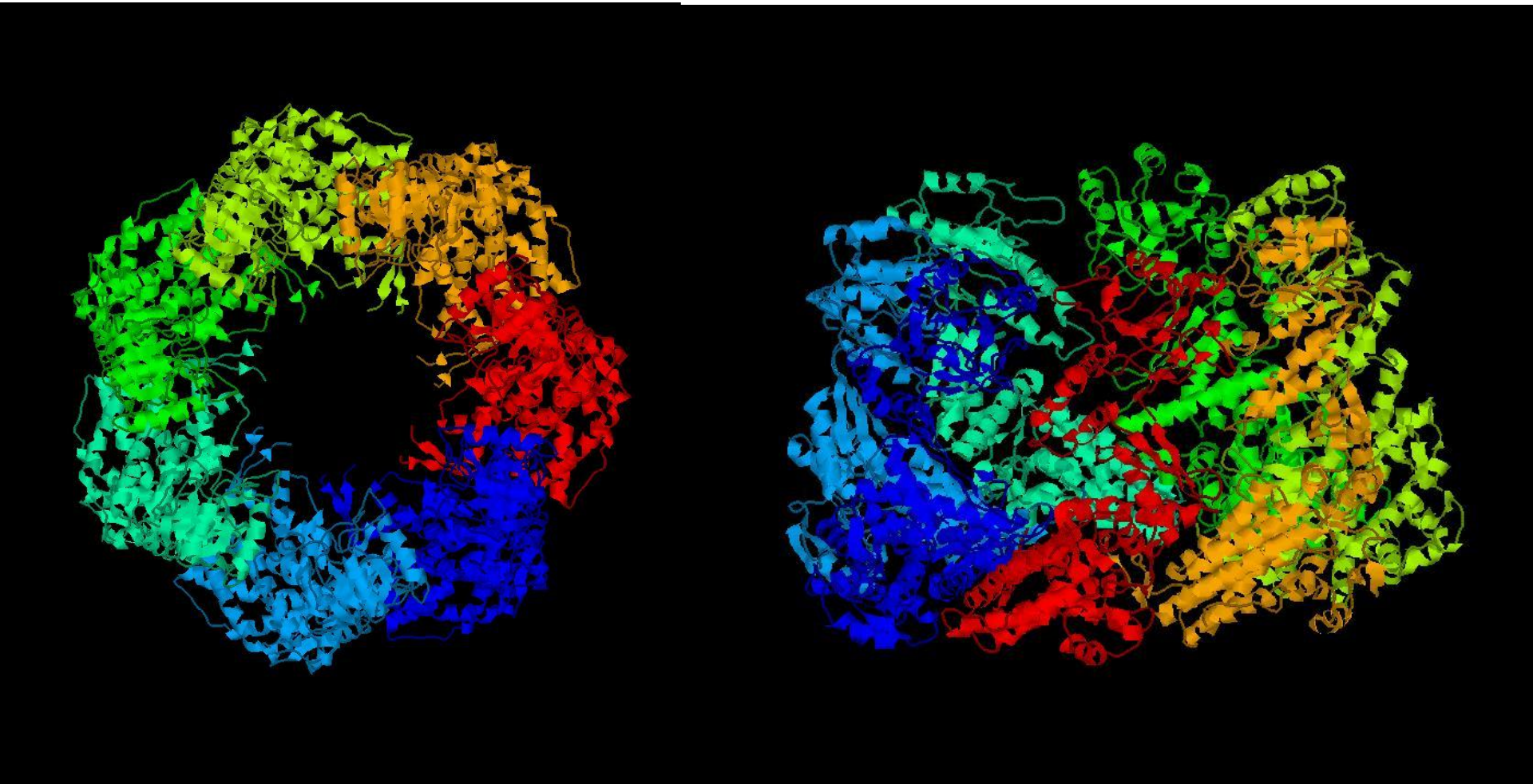
4 proteins

141-146-141-146 aa

GroEL (HSP)

top view

side view



http://www.rcsb.org/pdb/hor

RCSB Protein Data Bank - ...

Google 검색 공유 더보기 » Jooyoung Lee

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB Login

RCSB PDB PROTEIN DATA BANK

An Information Portal to 109822 Biological Macromolecular Structures

Search by PDB ID, author, macromolecule, sequence, or liga

Go

Advanced Search | Browse by Annotations

PDB-101 Worldwide PDB EMDataBank Nucleic Acid Database StructuralBiology Knowledgebase

Welcome

Deposit

Search

Visualize

Analyze

Download

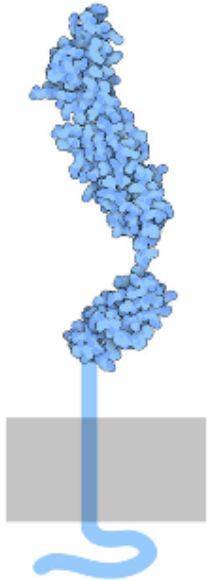
A Structural View of Biology

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

2015 High School Video Challenge Awards

June Molecule of the Month



What we are not sure about proteins:

- Secondary structure of a protein:
- Tertiary (3D) structure a protein
- Quaternary (4D) structure a protein

Protein folding problem

1. **Protein Structure Prediction:**
   For a given protein sequence, to determine its 3D structure by computation

2. **Protein-Folding Mechanisms:**
   By what process does a protein folds into its native and biologically active conformation?

3. **Inverse Folding:**
   For a given protein structure, to design its 1D sequence

What is CASP?

- **C**ritical **A**ssessment of Techniques for Protein **S**tructure **P**rediction (http://predictioncenter.gc.ucdavis.edu/).
- Goal is to help advance the methods of identifying protein structure from sequence.
- Community-wide experiments are held every two years starting 1994 (most recent one CASP10 in 2012)
- **Blind prediction and blind assessment**
- Since CASP1 (1994), there are a total of 758 protein sequences predicted.
- Since CASP5 (2002), ~200 methods have been tested for each CASP.

- There was a graduate student who knew nothing about quantum mechanics
- This poor student took a "quantum mechanics" course.
- He is about to take a take-home exam. He has two options to prepare the exam:

1. Examine last 90 year's problem sets **with answers**. → Homology modeling / Threading : template-based modeling → **"Multiple sequence alignment"**

2. Try to understand the problems and write down his own answers. → Ab initio (de novo, Energy-based,…) : physics-based modeling → **Global optimization of an accurate potential energy function**

Protein Structure Prediction

1. Physics-based approaches: Principle-based modeling
   ① Accurate potential energy function
   ② Powerful global optimization method → what we can do better than others
   ③ Ab initio, de novo, new fold targets (10-20%)

2. Informatics-based approaches: Template-based modeling
   ① Map the original problem to a problem with solution → mapping problem (alignment problem)
   ② Use templates (problems with solutions) to obtain the solution of the original problem (multiple alignment)
   ③ Comparative modeling, fold recognition (80-90%)

Protein Structure Prediction

1. Physics-based approaches: Principle-based modeling
   - ① Accurate potential energy function
   - ② **Powerful global optimization method**
   - ③ Ab initio, de novo, new fold targets (10-20%)

2. Informatics-based approaches: Template-based modeling
   - ① Map the original problem to a problem with solution → mapping problem (alignment problem)
   - ② Use templates (problems with solutions) to obtain the solution of the original problem (multiple alignment)
   - ③ Comparative modeling & fold recognition (80-90%)

Potential energy function: all-atom potential

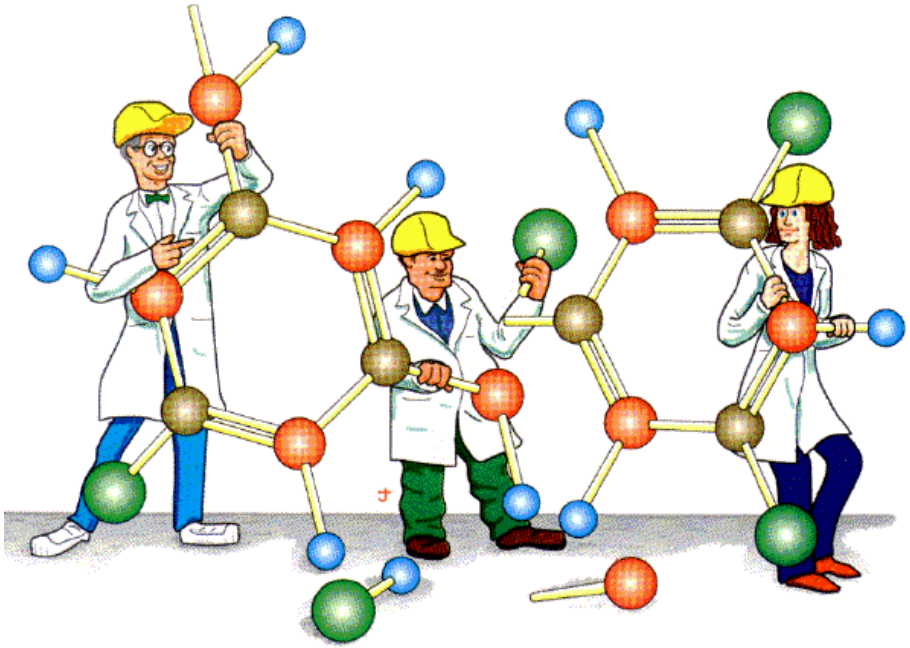
- All atom, off-lattice force field
- AMBER, CHARMM, GROMOS and others
- E terms: vibrational, torsional, non-bonded, electrostatic, and others including solvation

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

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

P2. TINKER Molecular Modeling

- Go to http://dasher.wustl.edu/tinker/ and download the TINKER package appropriate to your laptop.
- Perform Monte Carlo with Minimization to obtain the lowest energy structure of Met-enkephalin, a pentapeptide.
- 2 input files are provided here.
- Try 10 independent runs.
- What is the lowest energy?



Needs following 2 input files:

==> enkephalin.dat <==
enkephalin
Met-Enkephalin (YGGFM)
tyr
gly
gly
phe
met

n

==> enkephalin.key <==
parameters ../params/charmm22
maxiter 2000

randomseed 123456789

1. Generate an extended structure.
$ cat enkephalin.dat # review dat file
$ cat enkephalin.key # review key file

$ ../bin/protein.x < enkephalin.dat

$ cat enkephalin.seq # sequence file
$ cat enkephalin.xyz # tinker xyz file
$ cat enkephalin.int # internal coord

2. Energy component analysis
$ ../bin/analyze.xyz enkephalin.xyz e

3. To convert a xyz file to a pdb fiile
$ ../bin/xyzpdb.x enkephalin.xyz
$ cat enkephalin.pdb # pdb file

2. Monte Carlo with Minimization

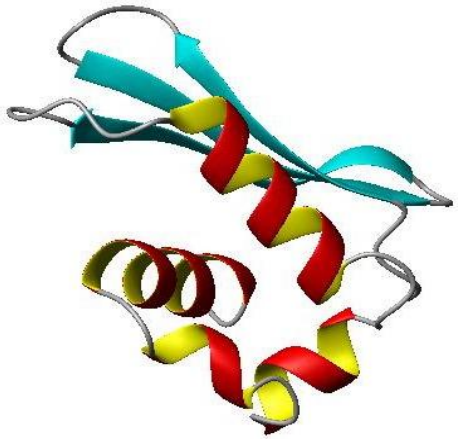
$ rm -rf enkephalin.xyz_2
$ ../bin/monte.x enkephalin.xyz

Past CASP Performances of KIAS protein folding lab

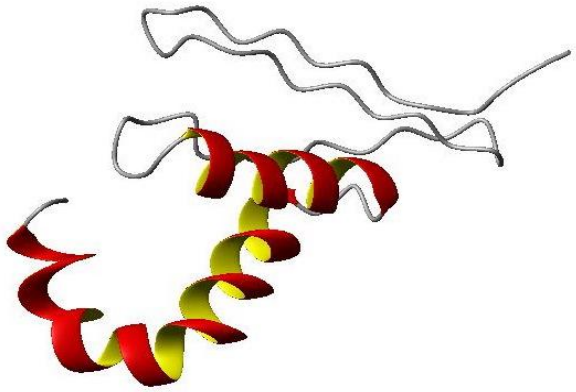
- CASP5 (2002): 18th out of 165 team in *new-fold* category
- CASP6 (2004): selected as a member of 12 elite teams in *new-fold*

CASP6 example: T0199_D3 (FR/A, Nres=82, 145-226)

Native structure



Model4



Physics & Protein Structure Prediction (I)

1. Proteins are polypeptide chains containing many (thousands of) atoms, and the interaction between atoms is considered to be reasonably well described by physics and chemistry.

2. However, there are only a few anecdotal examples of successful physics-based protein modeling (compared to the informatics-based method).

3. Currently, protein structure prediction methods relying only on physics-based approaches do not work as well as informatics-based methods.

Protein Structure Prediction

1. Physics-based approaches: Principal based modeling
   ① Accurate potential energy function
   ② Powerful global optimization method
   ③ Ab initio, de novo, new fold targets (10-20%)

2. **Informatics-based approaches: Template based modeling**
   ① **Map the original problem to a problem with solution → mapping problem (alignment problem)**
   ② **Use templates (problems with solutions) to obtain the solution of the original problem (*multiple alignment*)**
   ③ **Comparative modeling & fold recognition (80-90%)**

1.Sequence-sequence alignment

- How to align two given sequences.

  Possible alignments for KIAS and KAIST:

```
K-IAS-    KIA-S-
KAI-ST    K-AIST
```

2. Structure-structure alignment

- How to align two given sequences

3. Sequence-structure alignment

- Protein structure modeling

Pair-wise sequence alignment

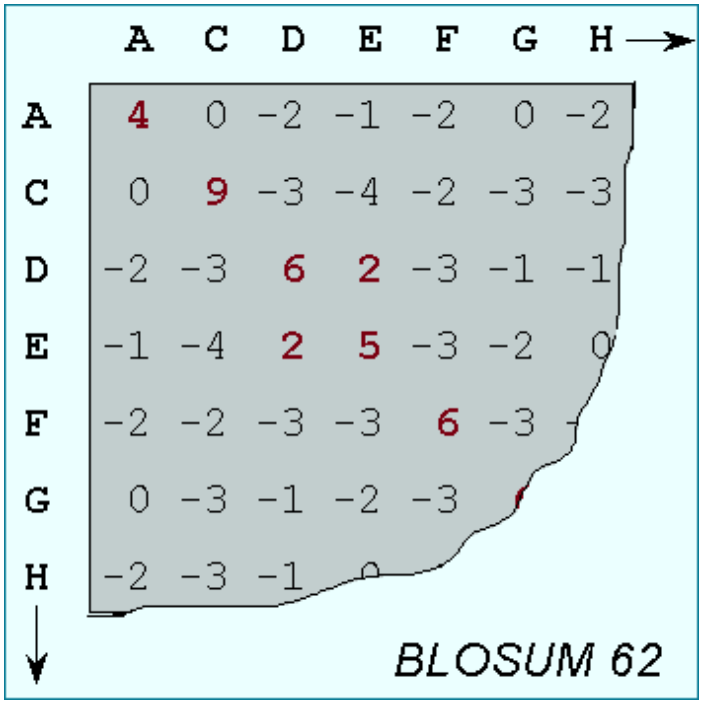
DP (dynamic programming) can provide exact results

```
Seq A: ARGTCAGATACGLAG---PGMCTETWV
Seq B: ARATCGGAT---IAGTIYPGMCTHTWV
```

Popular substitution matrices are PAM and BLOSUM.

$$S(A_L, B_L) = \sum_{i=1}^{L} sub(a_i, b_i) + G$$

$$G = G_{\text{open}} + G_{\text{ex}} \cdot n$$



Multiple Sequence Alignment (MSA)

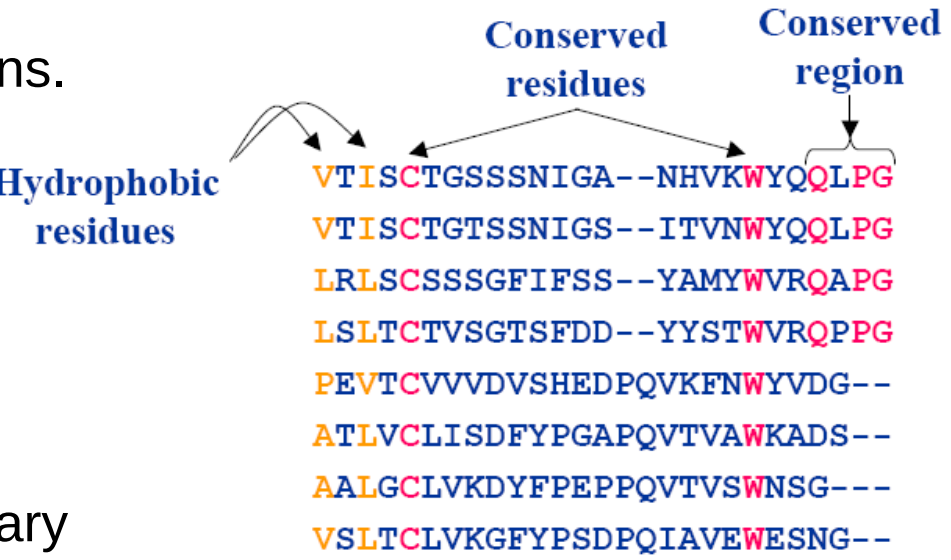
```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
             ***. ::: .: ..  .    :  . .      *  .  *: *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```

MSA → extension of pair-wise alignment.

MSA is useful for:

- Clustering, classification, or categorization of genes/proteins.
- Deducing evolutionary relationship and phylogenetic tree.
- Identification of conserved regions of genes/proteins .
- Detecting point mutations.
- predicting secondary and tertiary structure of proteins.



Sum of pair score

```
Seq A1: ARGTCAGATACGLAG---PGMCTETWV----
Seq A2: ARATCGGAT---IAGTIYPGMCTHTWVIAGQ
Seq A3: ARATCE--TACG--GTI-PGMCTHTWVIA--
```

$$Score(A_n) = \sum_{k=1}^{L} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sub(A_{i,k}, A_{j,k}) - \sum_{i=1}^{n} G(A_i) \qquad G(A_i) = a + bn$$

Exact method : multi-dimensional DP

-Time complexity O($L^n 2^n$), Space complexity O(L^n)

MSA objective function

COFFEE (Consistency based Objective Function For alignmEnt Evaluation) Score function: Given a set of sequences, the optimal MSA is defined as the one that agrees the most with all the possible optimal pair-wise alignments

$$COFFEE = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} W_{ij} \times Score(A_{ij})}{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} W_{ij} \times Len(A_{ij})}$$

-Score(A_{ij}) = Number of aligned pairs of residues that are shared between A_{ij} and the library.

-W_{ij} = Seq. Id btw sequences i,j

- do not depend on a specific substitution matrix
- the most consistent are often closer to the truth

> Construct pairwise alignment library based on profile-profile alignment

- Profile generate using PSI-BLAST (filter, -j 20 -h 0.0005)
- Alignment score : dot product score (Global-local DP, u=-1.2, v=-0.04, z=-0.05)
- Optimize : 630 pair of homstrad database

Physics & Protein Structure Prediction (II)

1. The goal is to achieve better protein modeling by **fusing informatics-based methods with a principle of physics (global optimization)**

2. The task was to map **protein modeling using templates** into a series of **combinatorial optimization problems**

3. The reality was **to learn TBM** (template-based modeling) **by making lots of mistakes in a real situation** (CASP7, 2006)

P3. Modeling a protein structure based on NMR restraints data

- Go to BMRB database http://www.bmrb.wisc.edu/ and download NOE and DIH restraints for 2G1E, or alternatively go to http://lee.kias.re.kr/~protein/wiki/doku.php?id=nmr:data : and download NOE and DIH restraints for 2G1E
- For a given correct distance pair, flat bottom restraint energy form can be used. That is for 1.8 < r < distance, no penalty is applied. But for r >distance, penalty in the harmonic form can be applied.
- Try to build a model of 2G1E which is consistent with the NMR data and all the stereochemistry of the protein (bond length, bond angle, no atomic clashes, etc)
- How similar is your model to the actual native structure of the protein?

P4. Modeling a protein structure based on ambiguous NMR data

- Go to the following page and search for Ts763: http://www.predictioncenter.org/casp11/targetlist.cgi
- Download the ambiguous NMR data of Ts763 (Ts763.tar.gz) from the following page and examine the data predictioncenter.org/download_area/CASP11/extra_experiments/
- Each line of the restraint data corresponds to an NMR peak arising from two hydrogen atoms positioned within a given distance. You should note that many peaks are represented by more than a distance pair, therefore the ambiguity arises. But, at least one of the provided distance pair is correct.
- For a correct distance pair, flat bottom restraint energy form can be used. That is for 1.8 < r < distance, no penalty is applied. But for r >distance, penalty in the harmonic form can be applied.
- Try to build a model of Ts763 which is consistent with the NMR data and all the stereochemistry of the protein (bond length, bond angle, no atomic clashes, etc)
- How similar is your model to the actual native structure of the protein?