Global Optimization by Conformational Space Annealing and its Applications to Biological Systems

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

MINI-WORKSHOP at JAPAN-KOREA HPC WINTER SCHOOL
Center for Computational Sciences, University of Tsukuba
Feb 26 , 2014

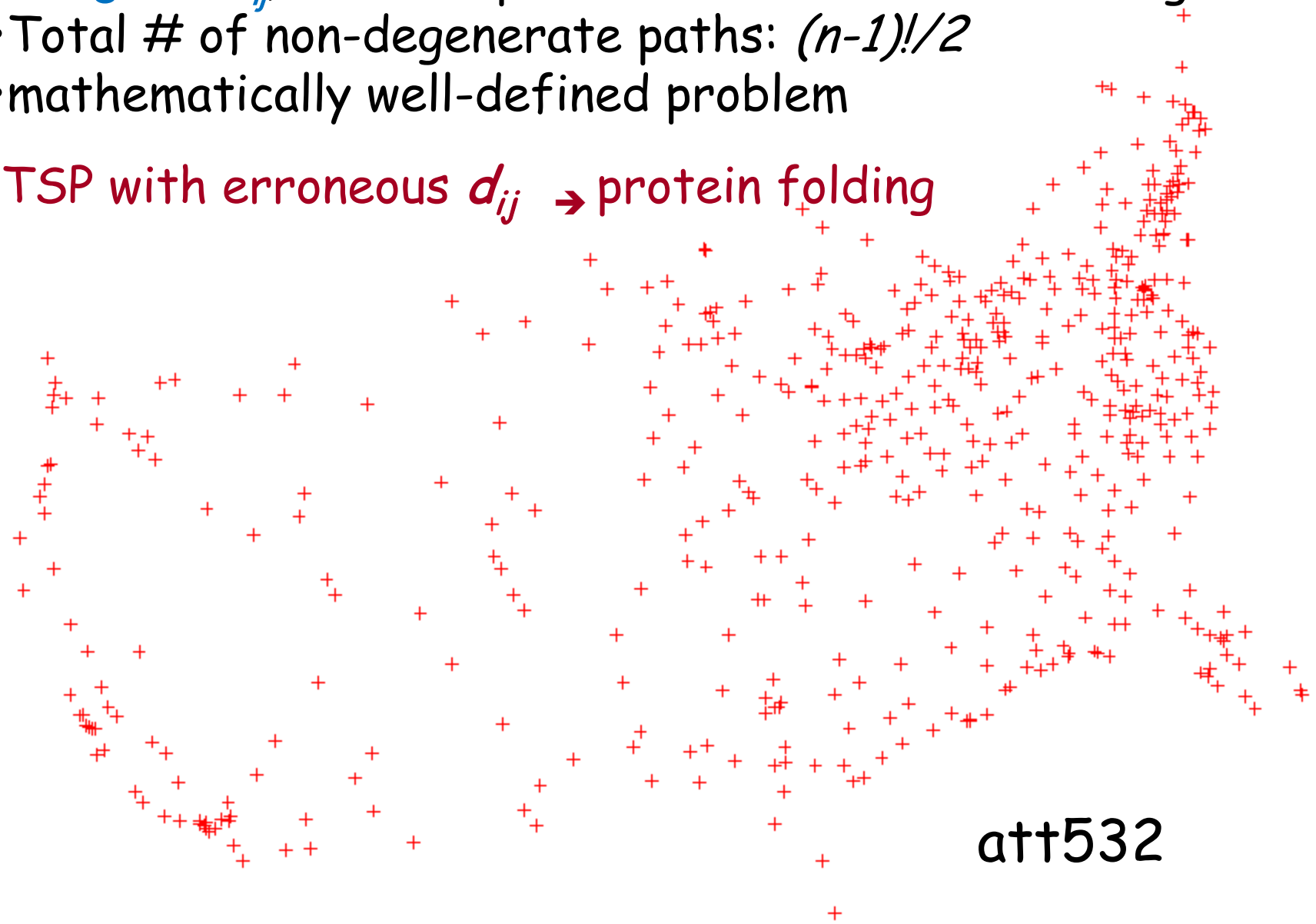
- **Protein Folding Problem**: Why is it so difficult to solve it by computation?
- Protein Structure 3D Modeling: Physics-Based vs. Informatics-Based
- High-Accuracy Protein Modeling by Global Optimization:
  -- CASP 7/8/9/10
- Accurate Protein 3D Modeling → Better Understanding of Biology?

Global Optimization

- Many problems in science and engineering are optimization problems.
- Efficient acquisition of the ground state and low-lying excitations is often sufficient to understand the essence of the problem.
- “Prediction” followed by “experimental validation” is one area where collaboration between theory and experiments can be most successful.
- Generation of more consistent models with experiments by directly optimizing a restraint function is another area where computation can contribute in many fields ➔ X-ray and NMR protein structure **“determination”.**

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



KIAS

e. kr

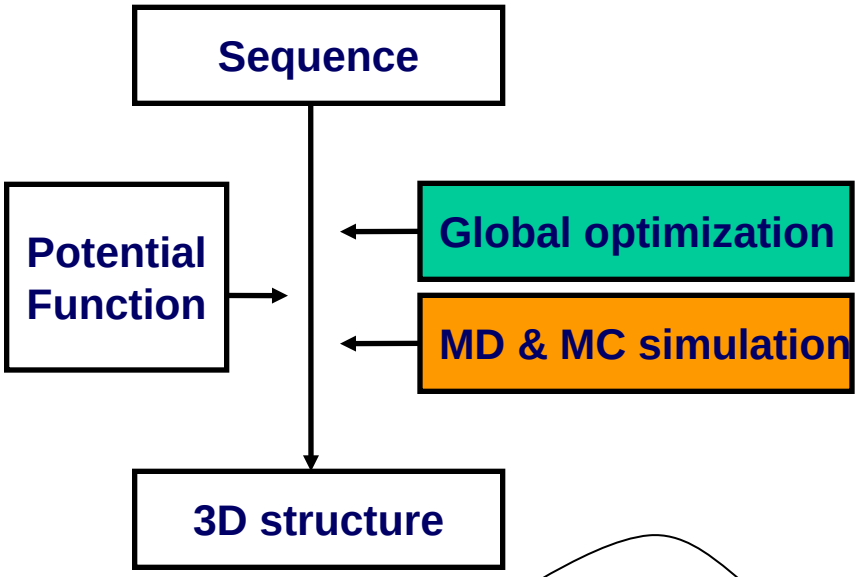
Ab initio protein folding by computer simulation

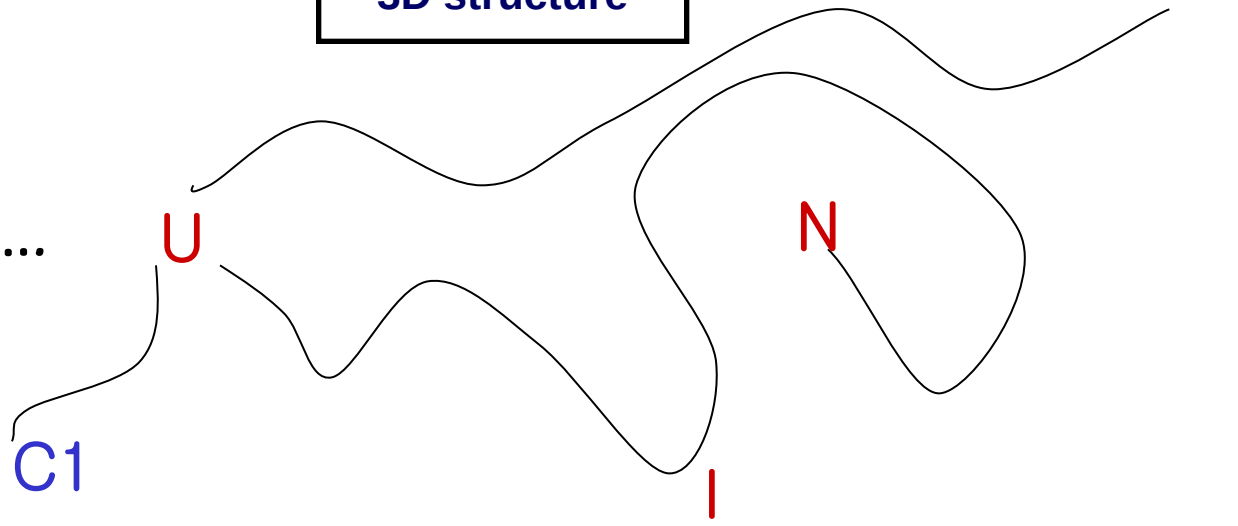
E(x)=…

E(x)=…

E(x)=…

E(x)=…





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%)**

Narrows the search

while maintaining diversity of sampling.

↓

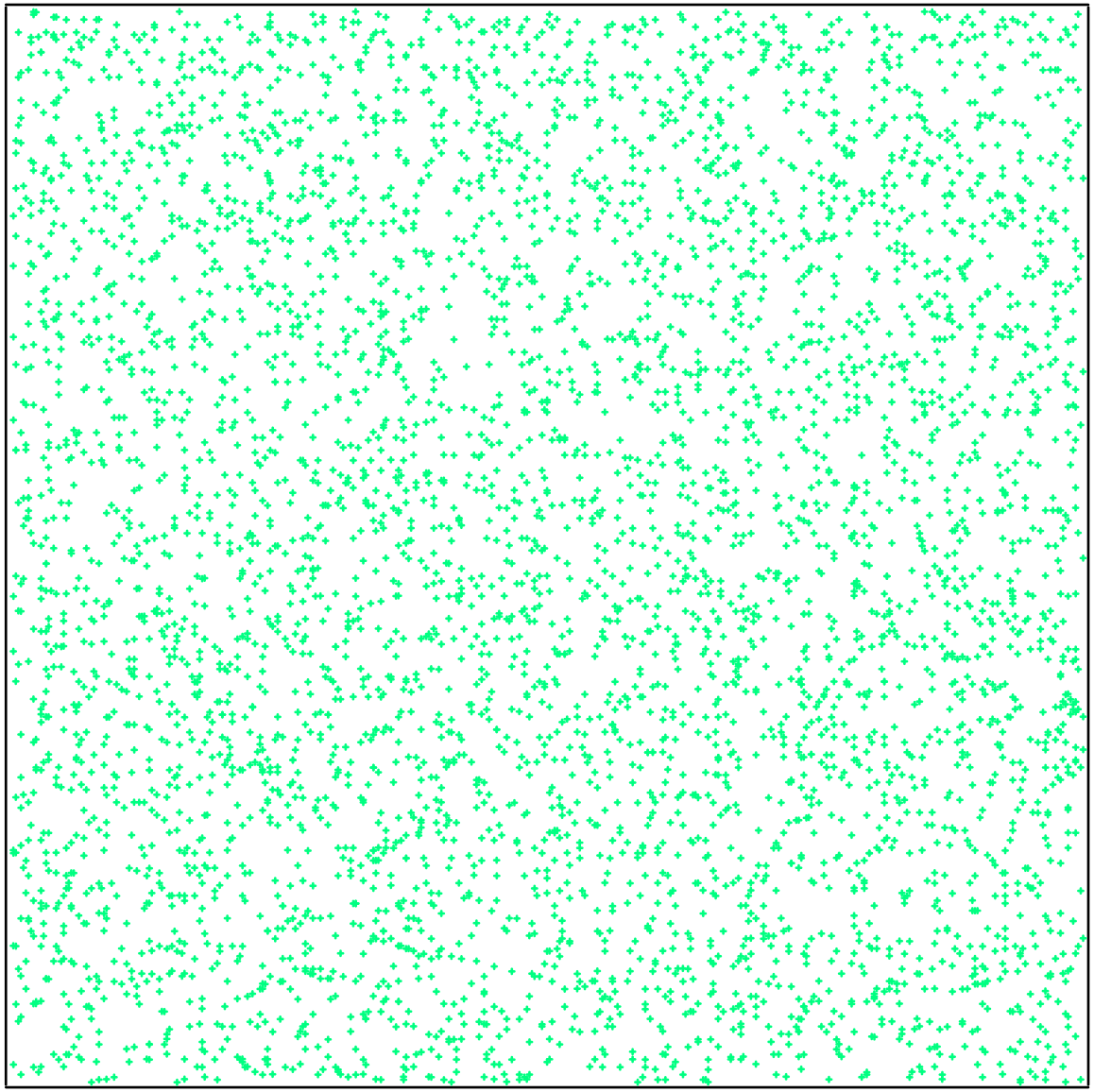
Annealing in conformational space".

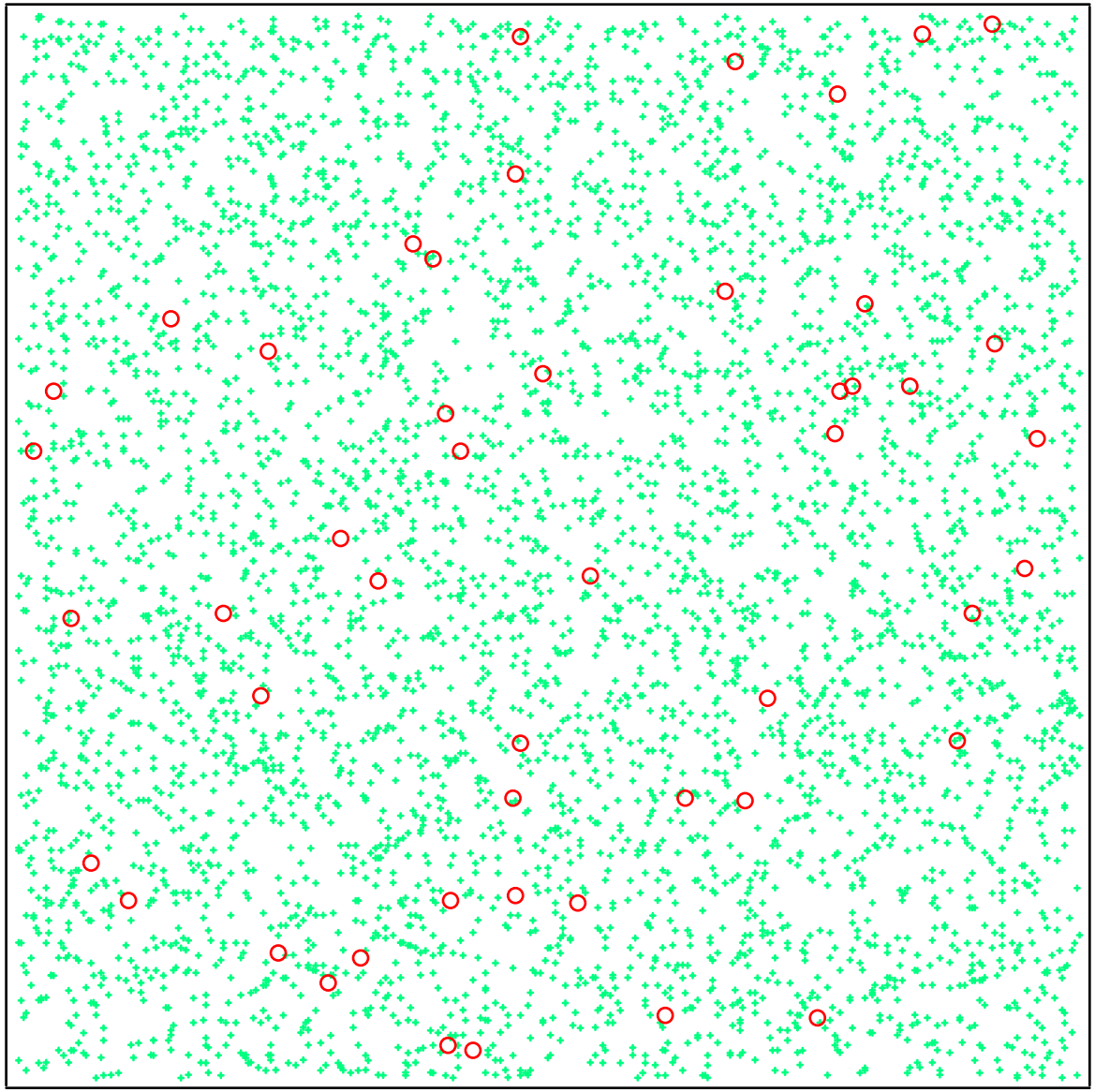
↓

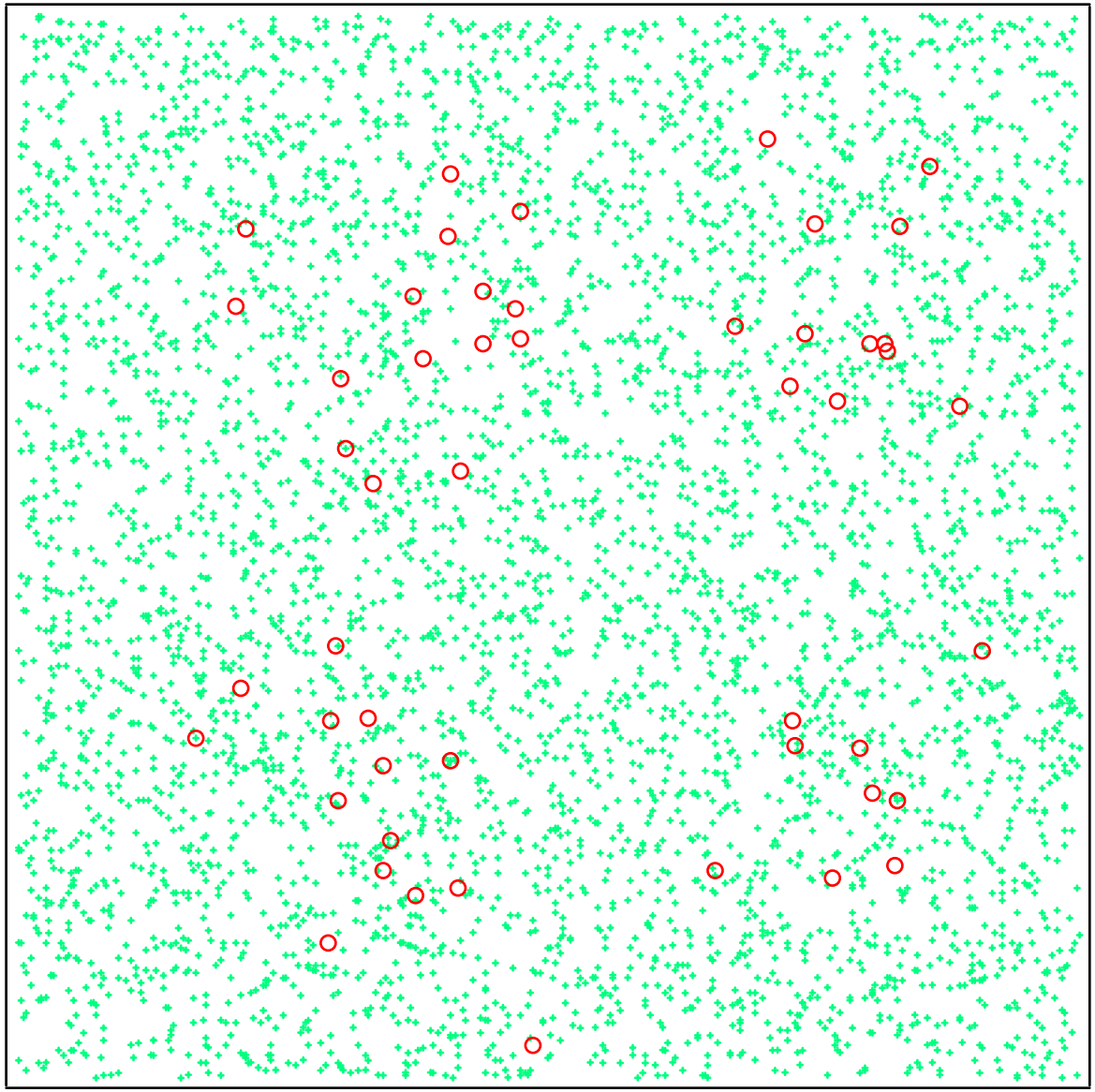
Conformational Space Annealing

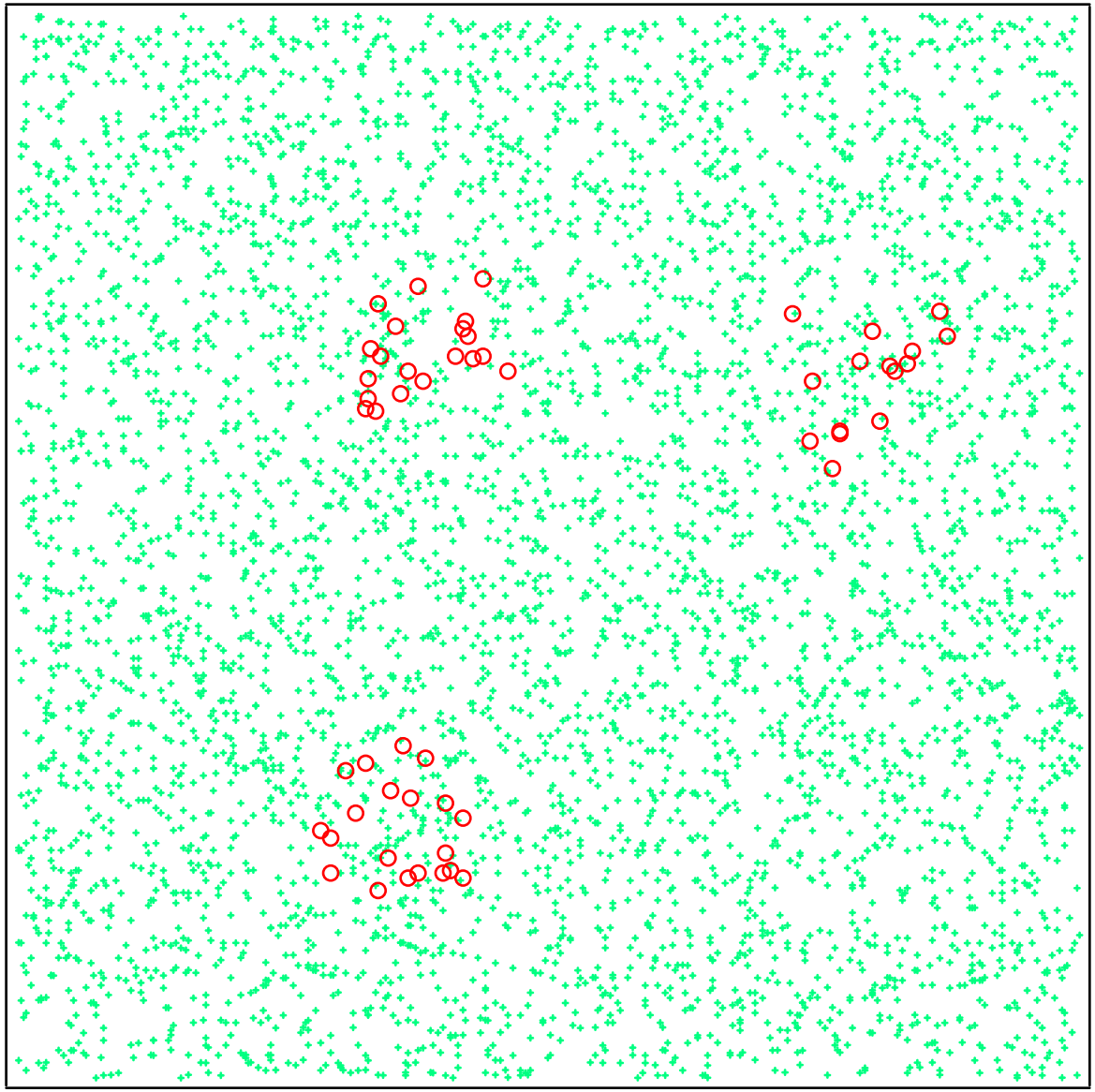
J Comput Chem **18** 1222 (1997)

Phys Rev Lett **91** 080201 (2003)

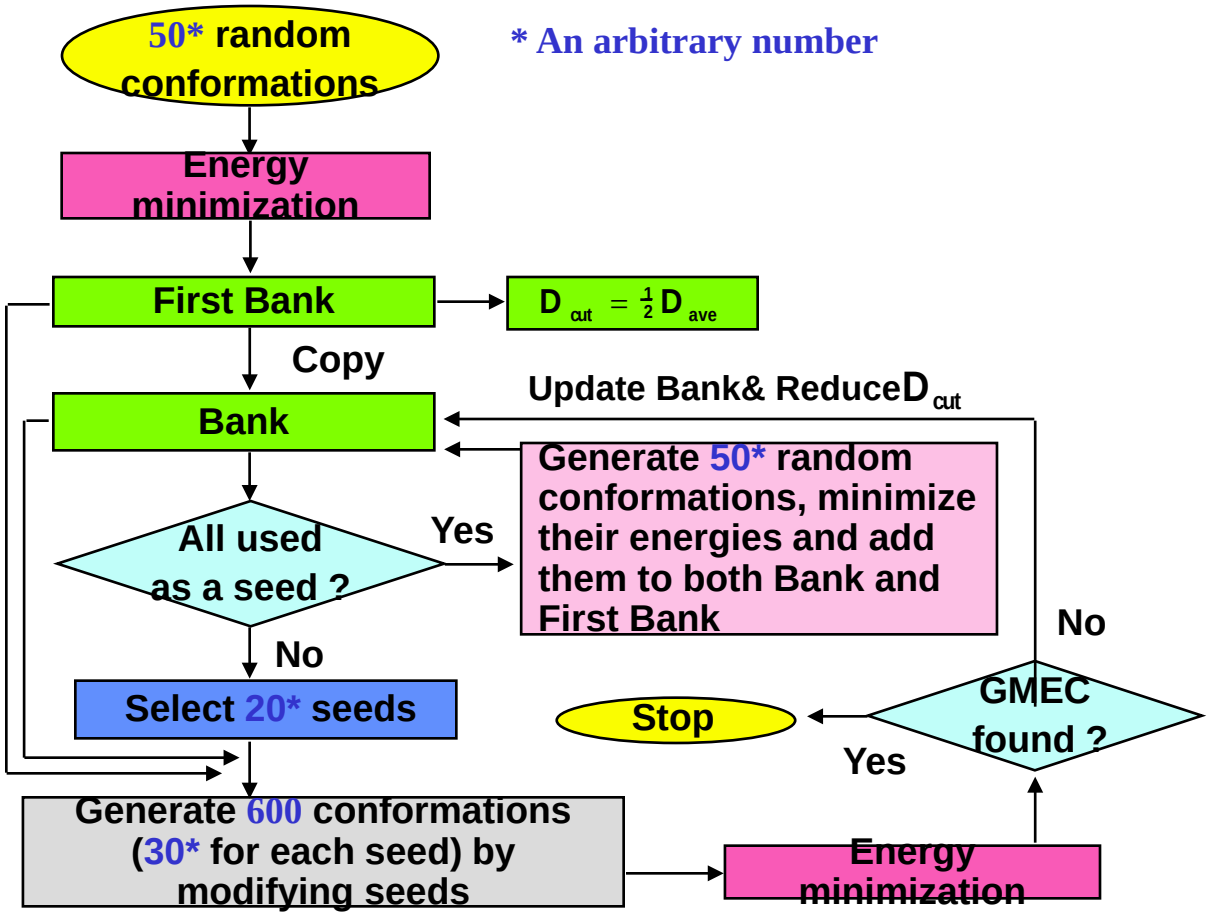


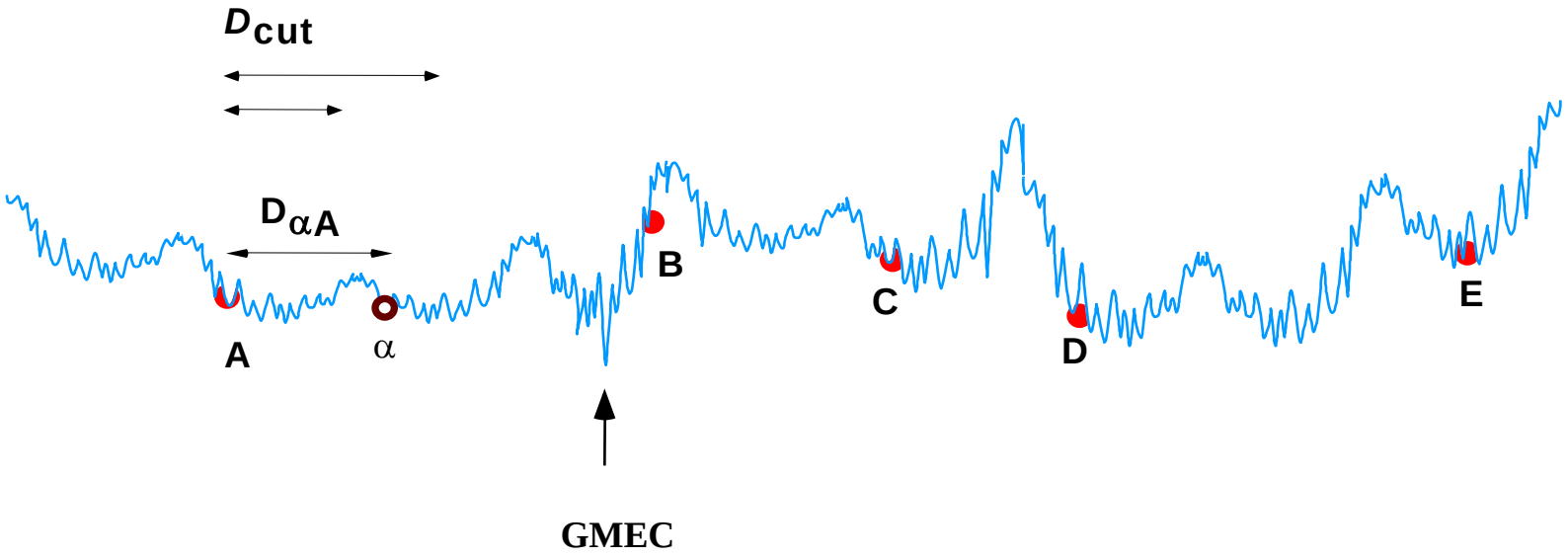






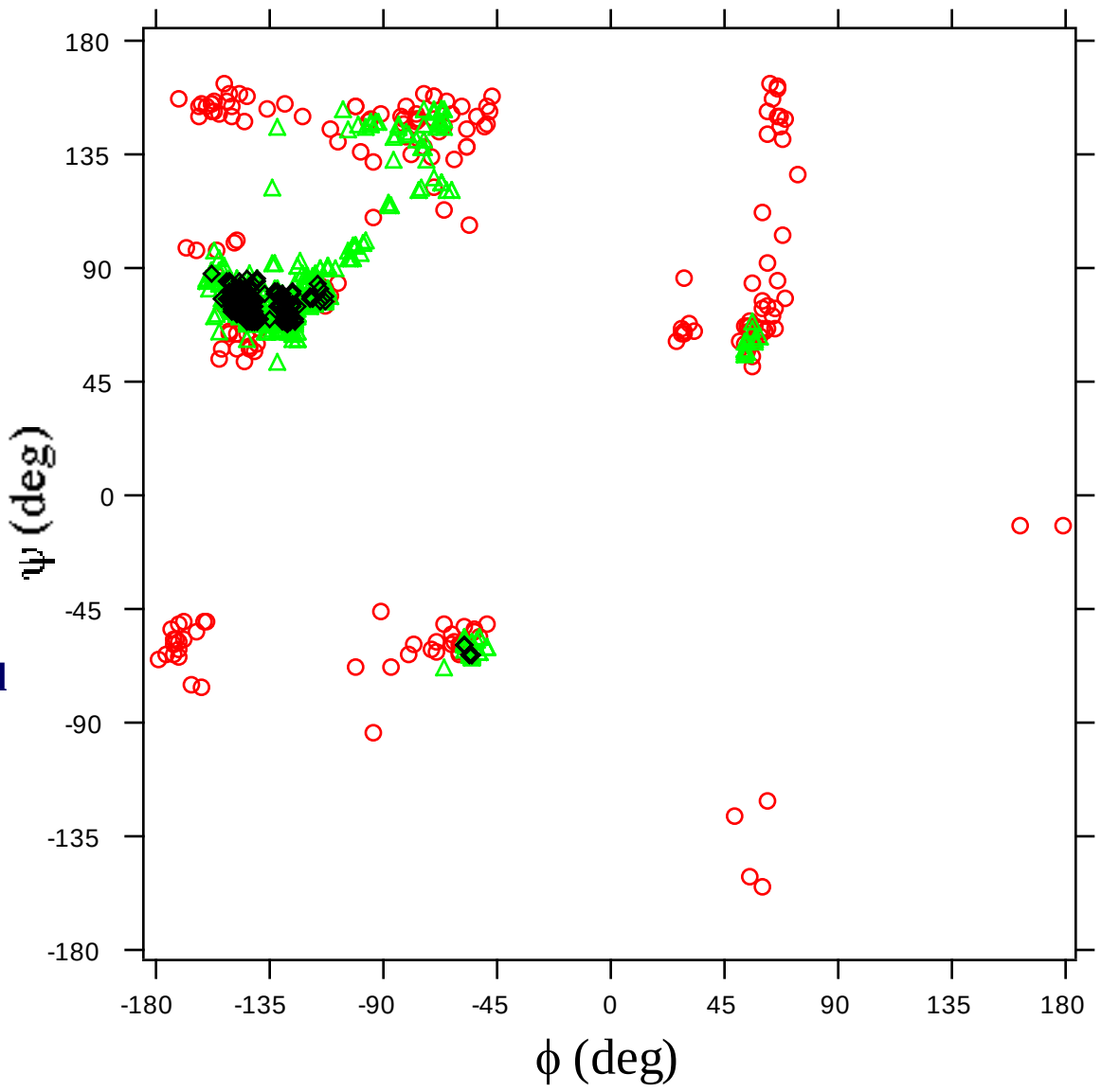
Flow chart of CSA





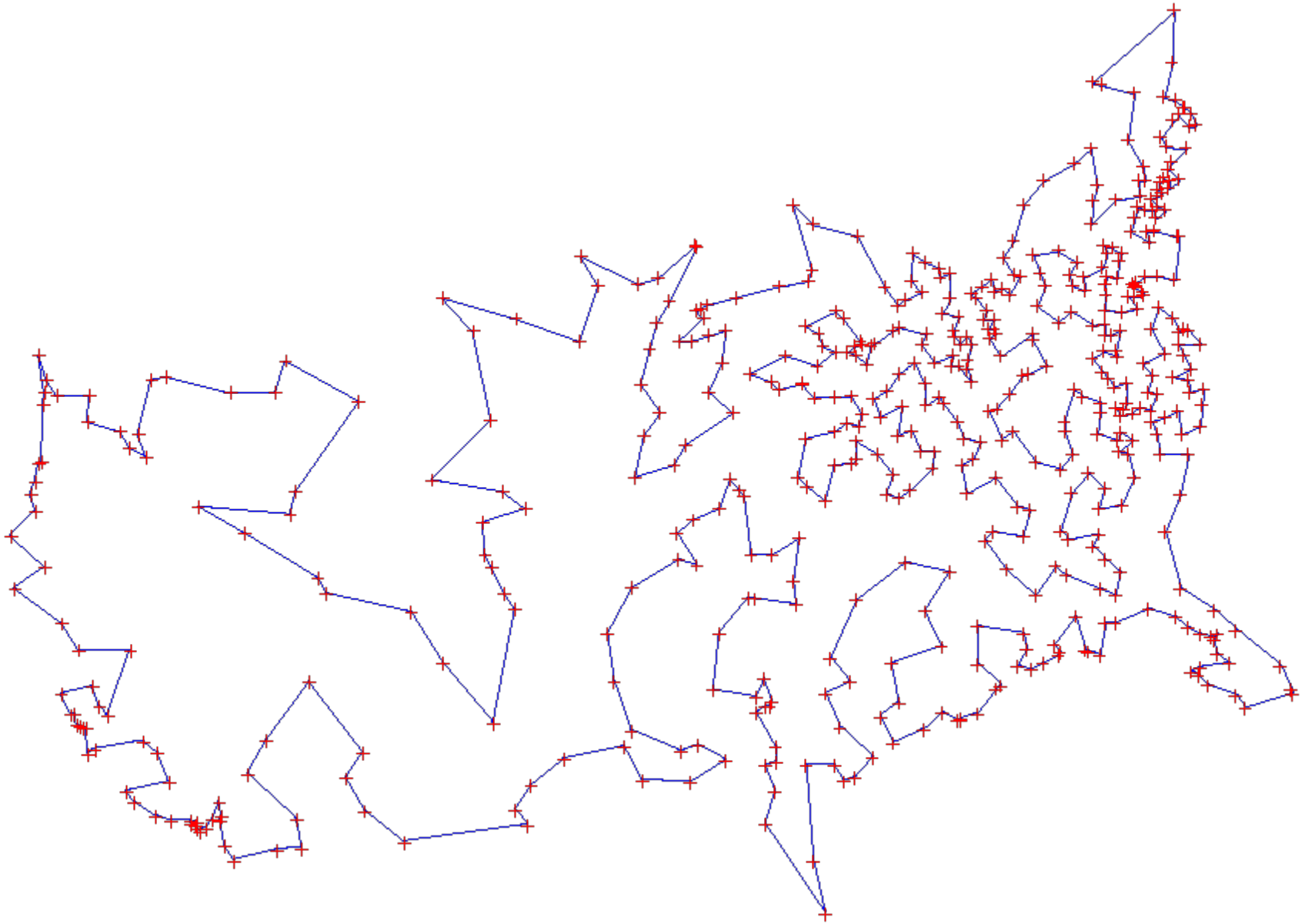
Schematic diagram to illustrate the procedure to update the bank conformations with a trial conformation α. The bank conformations (of size 5 in this example) are labeled with capital letters. First, by measuring all the distances between the trial conformation α and the bank conformations *A*, *B*, *C*, *D*, and *E*, find the closest conformation *A* (to α) at a distance $D_{\alpha A}$. The procedure to update the bank depends on the relative size of $D_{\alpha A}$ and D_{cut}. If $D_{\alpha A} < D_{cut}$ (with the larger size of D_{cut} shown in the figure) α replaces *A* if, in addition, α is lower in energy than *A*. However, if $D_{\alpha A} > D_{cut}$ (with the smaller size of D_{cut} shown in the figure) α replaces *B*, the highest energy (most unfit) conformation in the bank, if α, in addition, is lower in energy than *B*. If α does not satisfy the "lower in energy" condition in either of the two cases, α is discarded.

(ϕ, ψ) map of the Leu-13 residue of the 20-residue membrane-bound portion of melittin. In this peptide the Leu-13 residue precedes Pro-14. (113 variable angles)



Selected examples of successful optimization

- Optimization of ECEPP/3 for a 20-residue membrane-bound portion of melittin [Biopolymers **46**, 103-115 (1998)]
- Unbiased global optimization of Lennard Jones clusters up to N =201 [Phys Rev Lett **91**, 080201 (2003)]
- Ground state in the frustrated XY model and lattice coulomb gas with f =1⁄6 [Physica A **31**5 314-320 (2002)]
- Conformational space annealing and an off-lattice frustrated model protein [J Chem Phys **119** 10274-10279 (2003)]
- Structure optimization of an off-lattice AB protein model [Phys Rev E **72** 011916 (2005), Submitted]
- Efficient molecular docking using conformational space annealing [J Comput Chem **26** 78-87 (2005)]
- Ground-state energy and energy landscape of the Sherrington-Kirkpatrick spin glass [PRB **76**, 184412 (2007)]
- **Successful High-Accuracy Template-Based Modeling in the CASP7 experiments** [Proteins, **69**, 83-89 Suppl. 8 (2007)]
- Multiple sequence alignment by conformational space annealing [Biophysical J. **95** 4813-4819 (2008)]
- All-atom chain-building by optimizing MODELLER energy function using conformational space annealing [Proteins, **75**, 1010-1023 (2009)]
- LigDockCSA: Protein-Ligand Docking Using Conformational Space Annealing [J Comput Chem , **32**, 3226-3232 (2011)]
- Modularity optimization by conformational space annealing, [PRE, 85, 056702 (2012)]
- Hidden information revealed by optimal community structure from a protein-complex network improves protein function prediction [PLOS One (2013)]



att532

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.

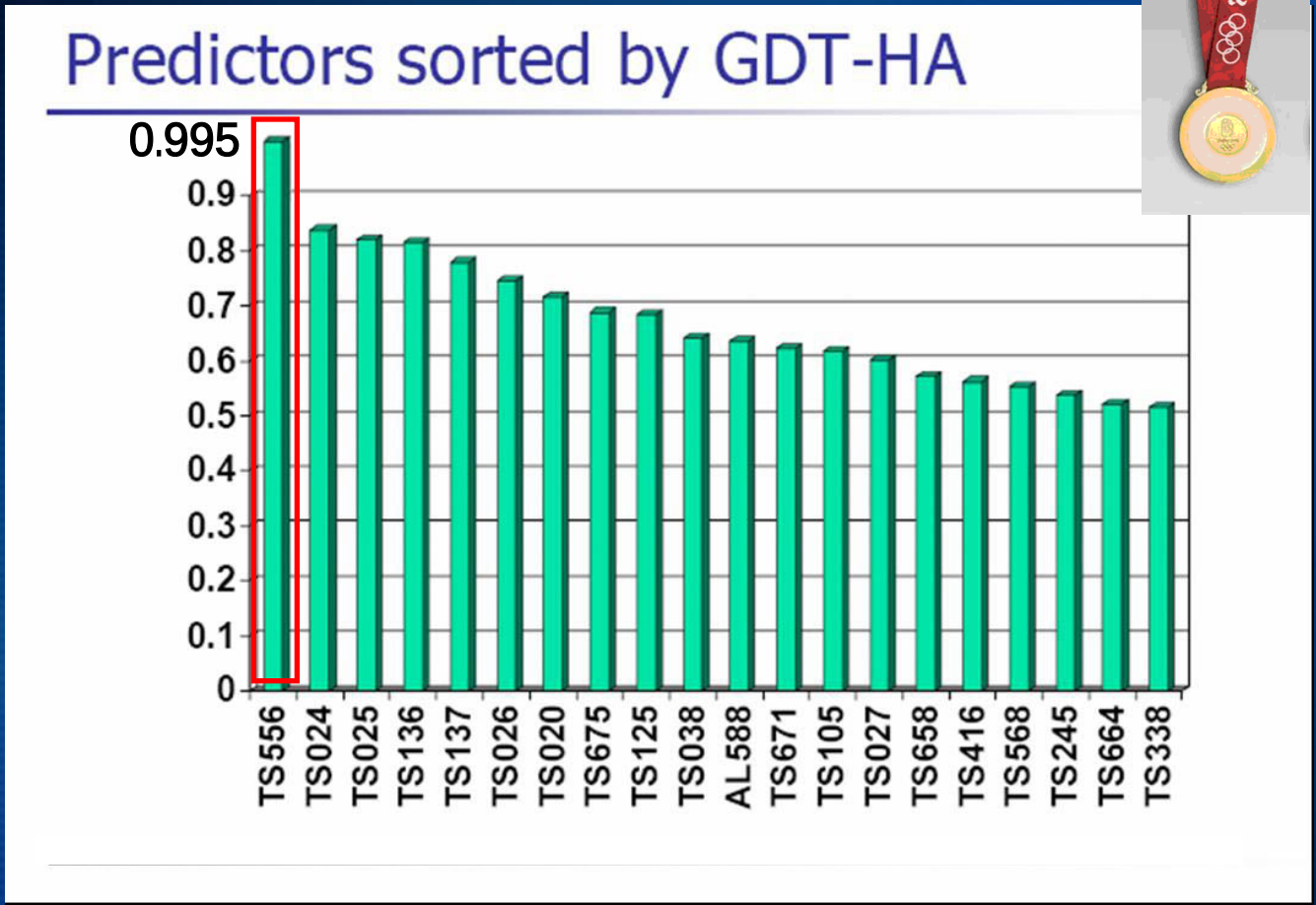
We formulate protein 3D modeling as a series of combinatorial optimization problems:

- **Multiple Sequence Alignment** (MSA) → optimization of a frustrate system [Biophysical J. **95** 4813-4819 (2008)]:
  - generate pair-wise alignments between all pairs
  - from each pair-wise alignment, generate residue-to-residue restraints → a library of restraints → a frustrated system
- **All-atom chain building** from MSA → another combinatorial problem of the modeller energy function [Proteins **75** 1010-1023 (2009)]:
  - modeller energy is a collection of competing terms including distance restraints from MSA and stereo-chemistry terms → inherent frustration when dealing with more than one template
  - modeller energy is treated as a black box for optimization
- **Side-chain modeling** is a combinatorial optimization of rotamers for a given backbone structure

CASP7 Experiment

- 2006, May -- August
- About 200 prediction methods are tested
- Total of 104 targets (9 cancelled)
- Three major categories:
  - **High Accuracy Template Based Modeling** (28 domains)
    - Use fine resolution measures for backbone assessment
    - Side-chains are also assessed
    - Only model 1s are considered
  - **Template Based Modeling** (108 domains)
  - **Free Modeling** (16 domains)
    - Physics-based methods have chances for providing competitive protein models
- Official results are available from CASP7 conference homepage (11/26-11/30/2006) and Proteins CASP7 issue

CASP7 High Accuracy Template Based Modeling



Proteins **69**, Issue S8, 27 – 37 (2007)

CASP7 High Accuracy Template Based Modeling

Group	n_{HA}	GDT-HA	AL0	1	1/2	n_{MR}	LLG	Sum
TS556 (LEE)	**26**	**0.995**	**0.727**	**1.427**	**1.290**	**12**	**0.842**	**3.127**
TS020 (Baker)	26	0.746	0.684	1.242	1.307	12	0.738	2.792
TS249 (taylor)	6	0.590	0.351	0.348	0.349	4	1.731	2.670
TS186 (CaspIta-FOX)	27	0.349	0.289	1.280	1.311	12	0.874	2.534
TS004 (ROBETTA)	28	0.432	0.382	1.405	1.290	12	0.792	2.515
TS671 (fams-multi)	28	0.654	0.657	0.876	0.933	12	0.616	2.203
TS010 (SAM-T06)	28	0.464	0.562	1.187	1.185	12	0.487	2.136
TS234 (McCormack-Okazaki)	2	0.414	0.338	0.865	0.672	2	1.028	2.115
TS664 (CIRCLE-FAMS)	28	0.588	0.630	0.907	0.924	12	0.510	2.022
TS209 (NanoDesign)	26	0.447	0.353	0.997	0.687	12	0.883	2.016
TS568 (CHIMERA)	28	0.574	0.636	0.768	0.752	12	0.688	2.015
TS559 (GSK-CCMM)	4	0.448	0.484	0.396	0.449	2	1.105	2.001
TS338 (UCB-SHI)	28	0.604	0.522	0.271	0.333	12	1.016	1.954
TS024 (Zhang)	28	0.838	0.795	0.561	0.679	12	0.411	1.928

A total of **174** groups

Proteins **69**, Issue S8, 27 – 37 (2007)

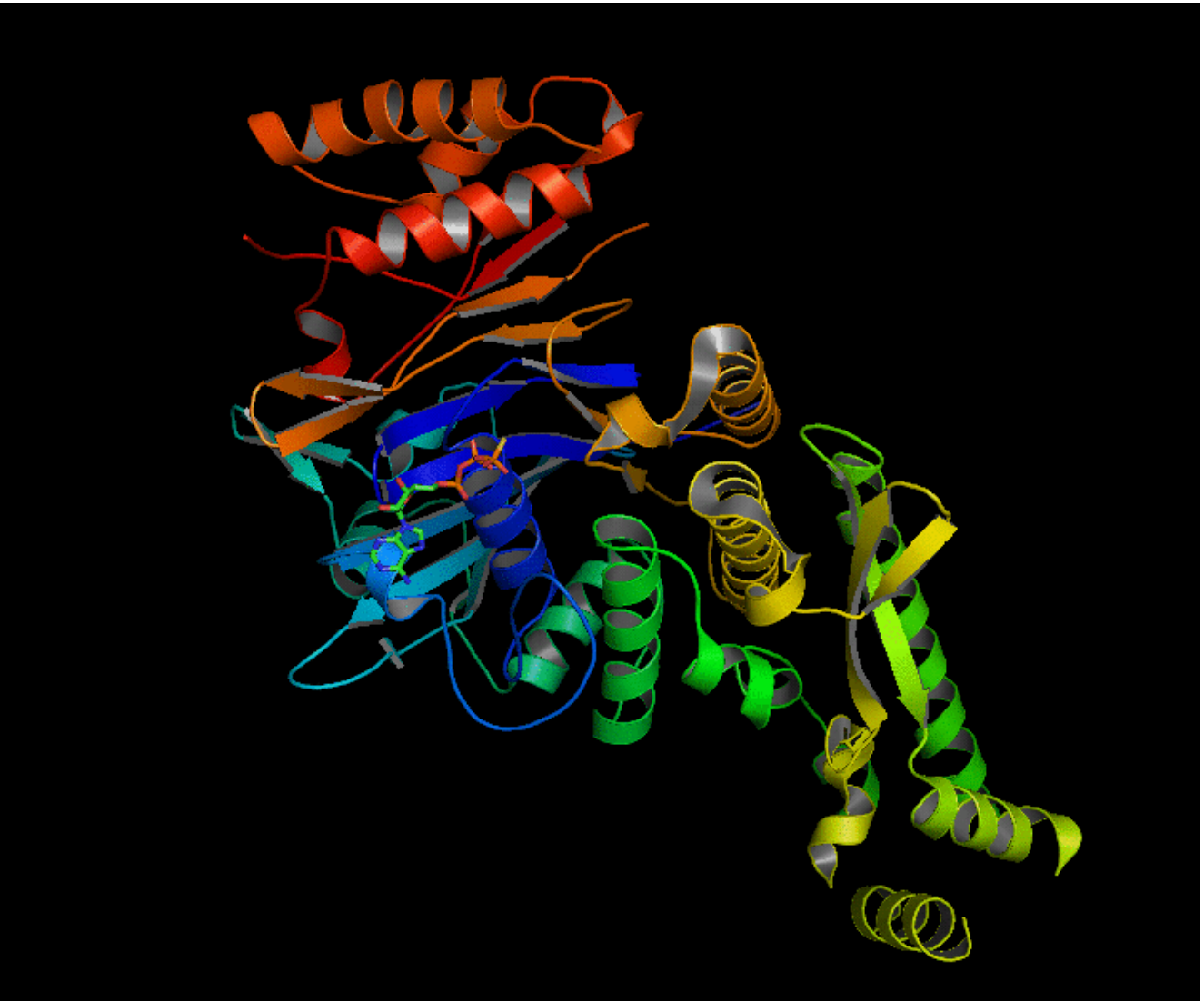
Conclusion of the official CASP7 assessment for HA/TBM targets [Proteins 69, Issue S8, 38 – 56 (2007)] reads:

"A number of groups did well in the HA/TBM category. ***Group 556 (LEE) stood out as the only group that performed near the top according to all criteria investigated***: fold quality (particularly GDT-HA), side-chain rotamer quality, and molecular replacement model quality".

Accurate protein models

Better understanding of biological mechanisms?

1. Determined a protein complex structure of condensin, MukBEF by combining X-ray data and protein modeling (with Prof BH Oh): “Structural Studies of a Bacterial Condensin Complex Reveal ATP-Dependent Disruption of Intersubunit Interactions” Cell **136** 85-96 (2009)



2. Screened natural proteins to find more efficient **w-aminotransferase** for asymmetric synthesis of chiral amine by protein modeling and docking simulation. The results are verified by wet experiments where 30-60 folds increased in the reaction rate is validated.
Biotechnology and Bioengineering **108** (2010) (with Prof BG. Kim)

Sequence name	Lowest distance among 100 docking poses (Å)	Initial rate for forward reaction (U_f, μmol/mg·min)	Initial rate for reverse reaction (U_r, μmol/mg·min)	U_r / U_f	Produced (*S*)-α-MBA (mM)
Atu4761	2.87	0.38	0.24		14.5
SAV2612	3.00	0.32	0.35		15.5
Atu3407	3.09	0.088	0.058		6.15
ω-ATV*f*	3.21	3.4	0.062	0.018	9.13

1. Caulobacter w-TA were selected and PSI-BLAST was run: 250 sequences were selected.
2. 250 sequences were multiply-aligned and 4 subgroups were identified.
3. 51 sequences belong to w-TA and all the sequences were used for model building.
4. The models were docked with aminodiphenylmethane(ADPM), and the distance between PLP and the N atom of ADPM was measured.

"Community/Module Detection" by Modularity Optimization

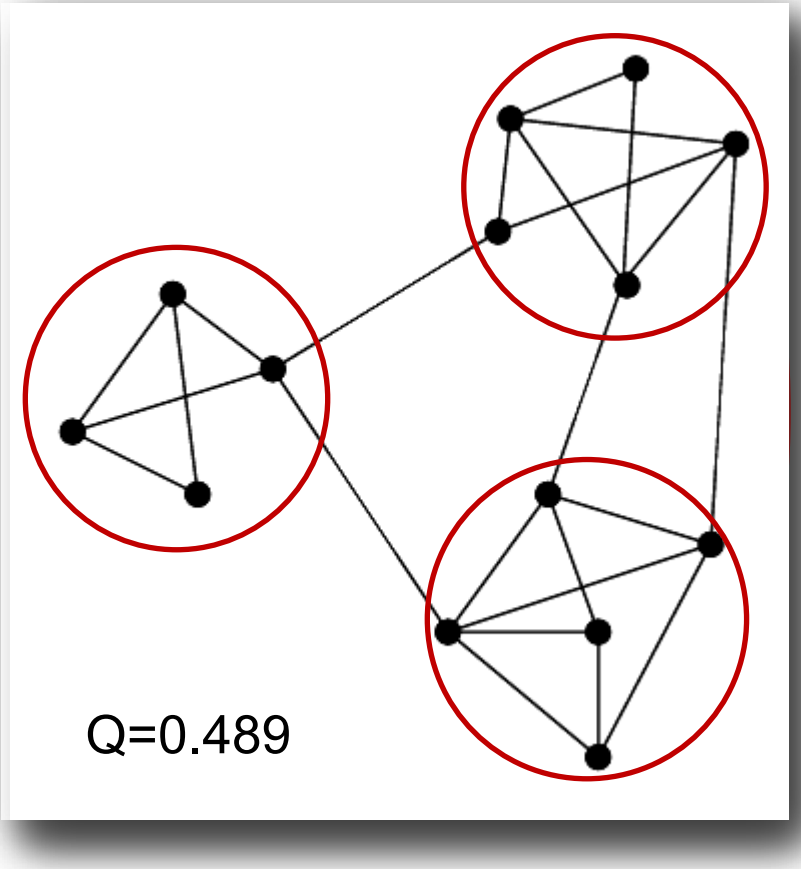
- Divide a network into sub-graphs/modules
  - nodes are more densely connected internally
- The most commonly used objective function to evaluate the quality of partition is Q proposed by Girvan and Newman

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

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



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.

Conclusions

- We have successfully mapped the **template-based protein modeling** into **three layers of combinatorial optimization problems**: **MSACSA, ModellerCSA** and **ROTCSA**.
- We have demonstrated that high accuracy protein 3D modeling can be achieved **simply by rigorous and straightforward optimization of score functions**.
- The proposed method requires a large amount of computational resources (100 CPU days per 300aa protein), but produces significantly better results.
- There are rooms for improvement by **better template detection** and **loop modeling**
- Application to real/experimental systems is in its preliminary stage but quite promising.

Acknowledgements

3D Modeling: Keehyoung Joo, *KIAS*
Jinwoo Lee, *Kwangwoon U.*
Sung Jong Lee, *Suwon U.*

Protein Function: Mina Oh, *KIAS*

Community Detection: Juyong Lee (KIAS/NIH)
Steven Gross (UC Irvine)

Experimental Collaboration: Byung-Gee Kim, *Seoul National U.* Byung-Ha Oh, *KAIST.*
DH Shin, *Ewha Womans U.* HC Shin, *Soongsil U.*
Weontae Lee, *Yonsei U*

International Collaboration: Masaki Sasai, *Nagoya U.* Bernie Brooks, *NIH*
Adam Liwo and Cezary Czaplewski, *U of Gdansk*
Zengyi Chang, *Beijing U*

Cluster Computers: *KIAS/CAC*

Supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korean government (MEST) (No. 2009-0063610)

Postdoc/Researcher Positions Available


Thank You!