1.Sequence-sequence alignment

- How to align two given sequences.

  Possible alignments for KIAS and KAIST:

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

2. Structure-structure alignment

- How to align two given sequences

3. Sequence-structure alignment

- **Protein structure modeling**

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)

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

Seq A_1: AR**G**TC**A**GAT**ACGL**AG---PGMCT**E**TWV----

Seq A_2: AR**A**TC**G**GAT---**I**AGTIYPGMCT**H**TWVIAGQ

Seq A_3: ARATCE--TACG--GTI-PGMCTHTWVIA--

The system is intrisically frustrated as in the SK spin-glass system.

$$\mathcal{H} = -\Sigma_{\langle ij \rangle} \frac{J_{ij}}{\sqrt{N}} S_i S_j$$

3D Modeling by Global Optimization

- Energy Function

$$E = E_{templates} + E_{stereo-chemistry} + E_{vdw}^{replusive} + E_{DFA} + E_{dfire} + E_{goap} + E_{hbond}$$

$E_{template}$: restraints from templates (Lorentzian shape)

E_{DFA}	: dynamic fragment assembly term
E_{dfire}	: dfire statistical potential term
E_{goap}	: orientation-dependent statistical potential term
E_{hbond}	: local hydrogen bonding term

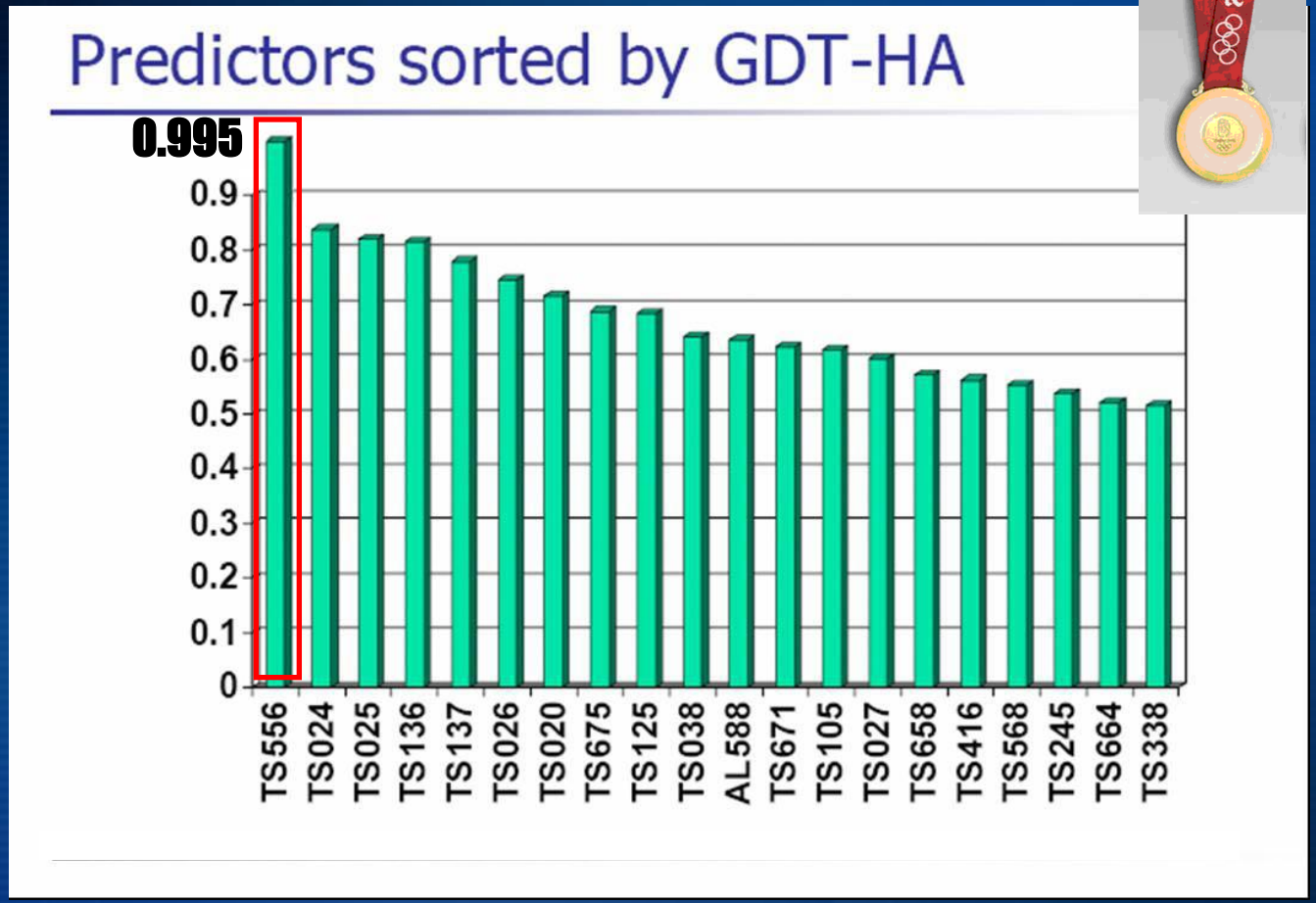
ab initio modeling terms

- Global Optimization by Conformational Space Annealing (CSA)

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".

CASP 7--10

- May-Aug of 2006, 2008, 2010 and 2012.
- For each CASP, over 200 prediction methods are tested.
- We tried 2 methods: LEE and LEE-SERVER (server)
- **Highlights** of LEE & LEE-SERVER predictions:
  - For TBM targets:
    - C^{α}+ other details: LEE & LEE-SERVER are top methods of choice.
    - Models are good not only in backbone accuracy as well as side-chain accuracy.

- CASP11 was carried out during May-Aug of 2014, and the result is available from the CASP11 webpage
  - LEER/nns is one of the top 5 methods for FM targets.
  - LEER is the best method for TBM targets.
  - LEER/LEE/nns is the best method for distance-information-assisted targets intended for solving large proteins using NMR spectroscopy.

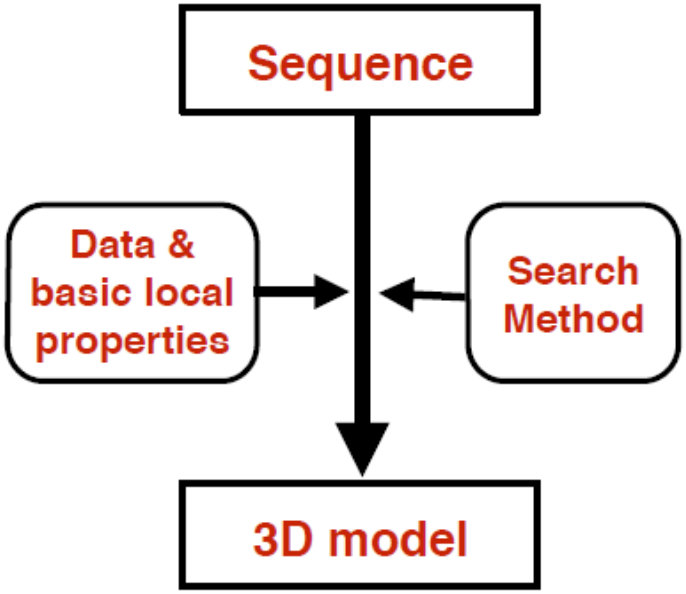
CASP11 report

(1) Template–free modeling of proteins
(2) Template–based modeling of proteins
(3) Protein structure modeling using sparse & ambiguous NOE restraints

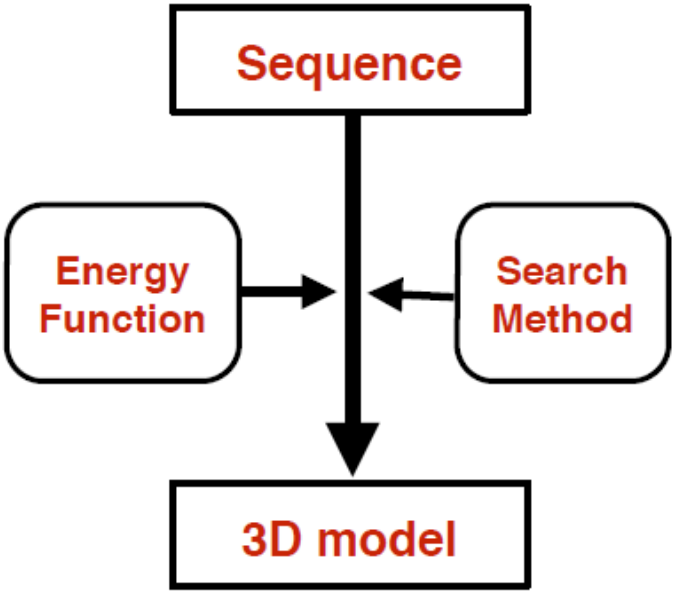
Dec. 8, 2014
Iberostar Paraiso, Riviera Maya, Mexico

- Fold Recognition: Sung Jong Lee (U. of Suwon, Korea) & Keehyoung Joo
- Protein 3D Modeling: Keehyoung Joo, InSuk Joung, & Sun Young Lee
- Model Refinement: InSuk Joung & Qianyi Cheng
- Quality Assessment: Sun Young Lee & Balachandran Manavalan
- Database: Jong Yun Kim
- Community Detection and X-ray crystal B-factor: Juyong Lee (NIH, US)
- Others: Jong Young Joung, Seungryong Heo, Mikyung Nam, In-Ho Lee (KRISS, Korea)

Data-driven modeling (determination/T[sc])



Strategy-dependent modeling (prediction/T[0px])



*Search Method: Conformational Space Annealing (CSA)

Energy/score function contains:
physics/statistical/bioinformatics terms

3D Modeling by Global Optimization

- Energy Function

$$E = E_{templates} + E_{stereo-chemistry} + E_{vdw}^{replusive} + E_{DFA} + E_{dfire} + E_{goap} + E_{hbond}$$

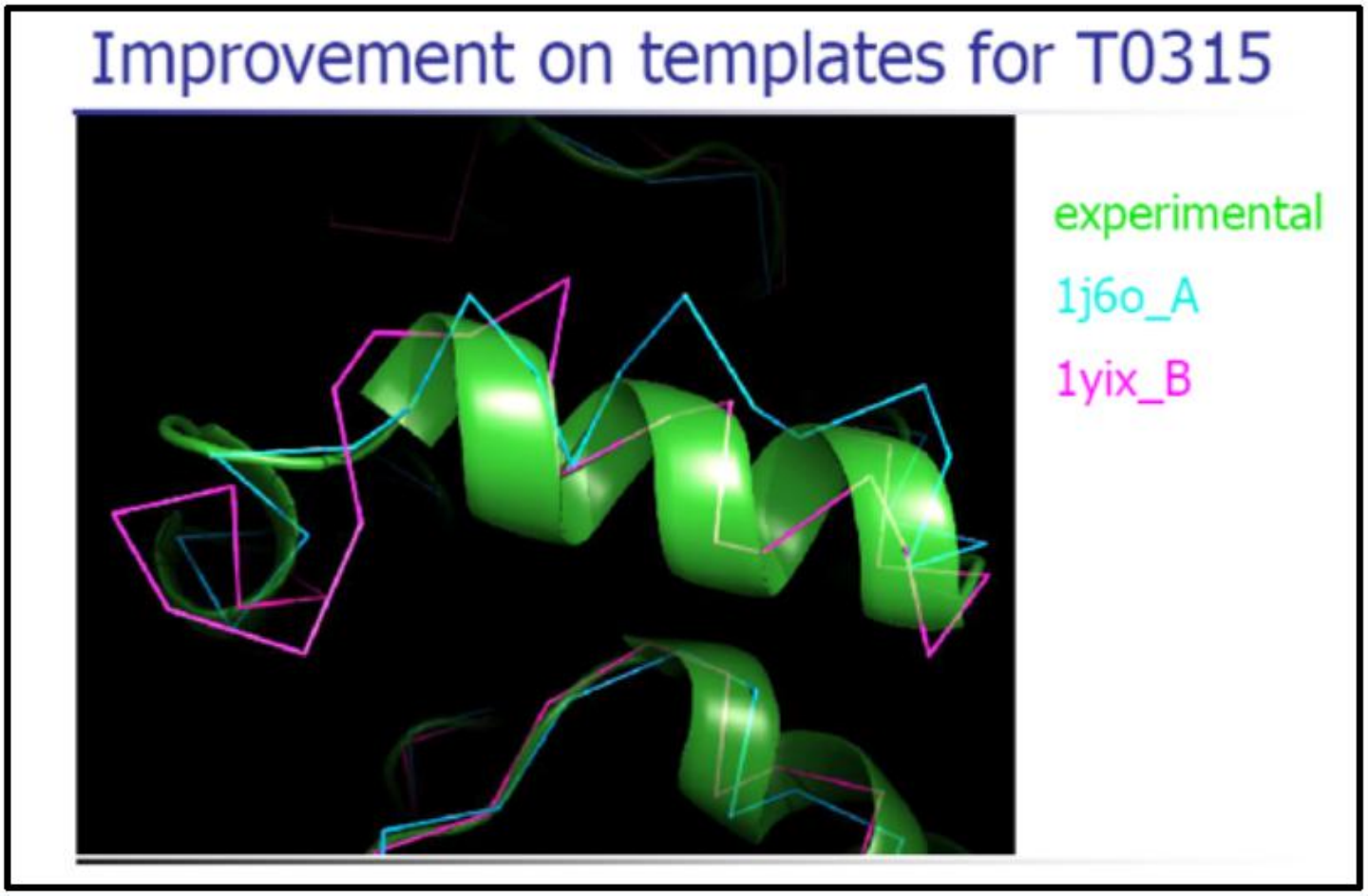
$E_{template}$: restraints from templates (Lorentzian shape)

E_{DFA}	: dynamic fragment assembly term
E_{dfire}	: dfire statistical potential term
E_{goap}	: orientation-dependent statistical potential term
E_{hbond}	: local hydrogen bonding term

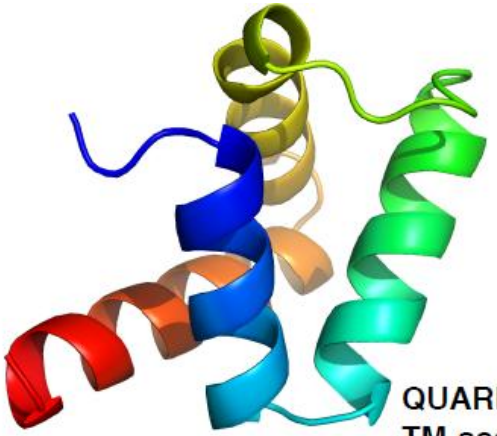
ab initio modeling terms

- Global Optimization by Conformational Space Annealing (CSA)

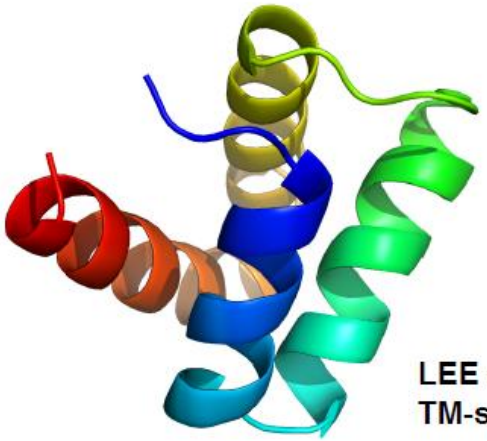
CASP7 Assessor's presentation by Randy Read for High-Accuracy TBM



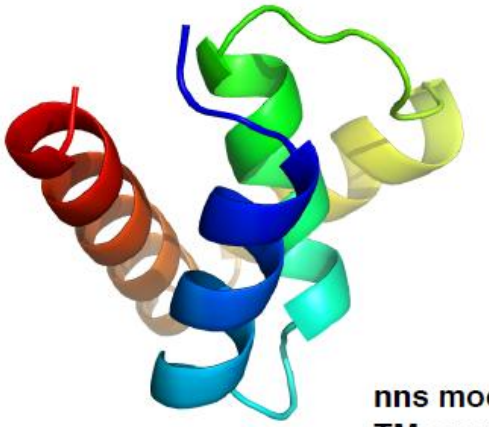
The challenge is to extract max information from multiple templates without the native structure information.



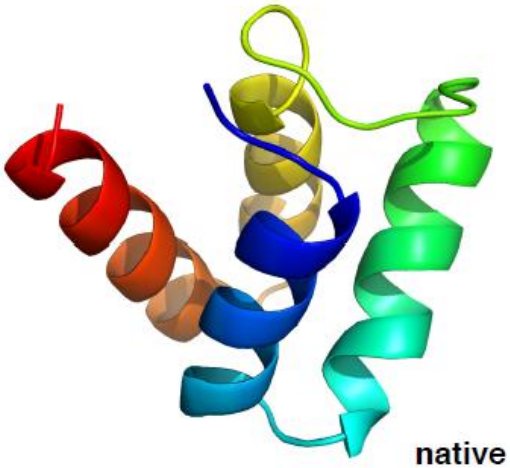
QUARK model1
TM-score = 0.5654



LEE model1
TM-score = 0.6740

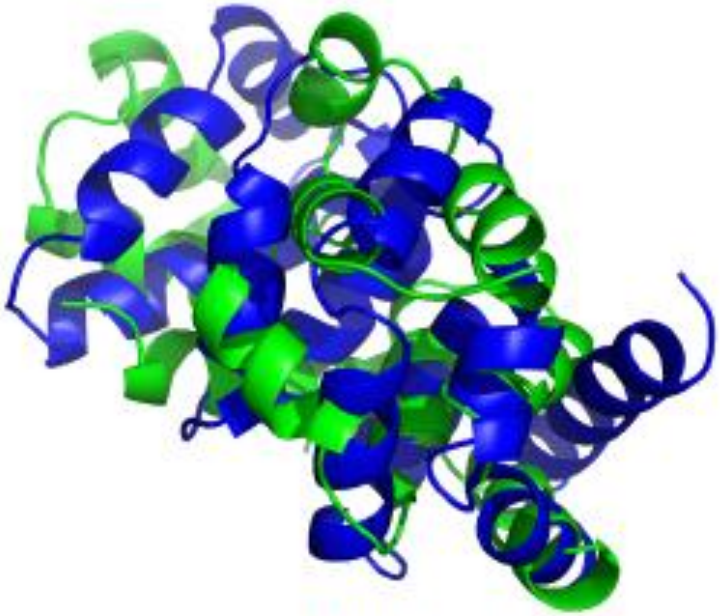


nns model1
TM-score = 0.4392

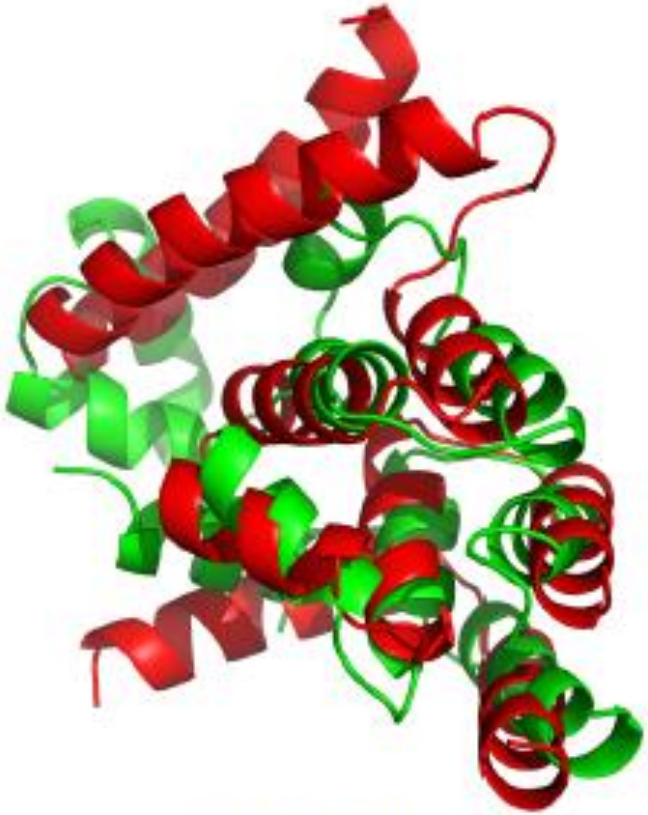


native

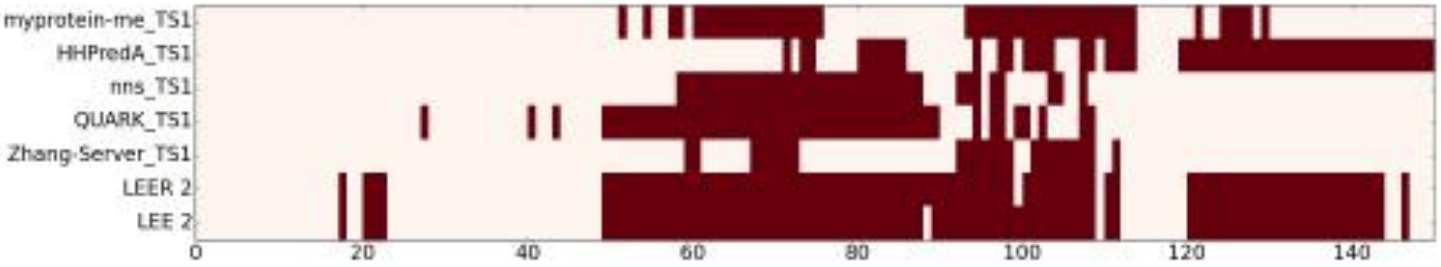
LEE model1 for T0816-D1 was generated using 12 templates with TM-scores ranging from 0.29 to 0.56 including two models on the left. In the LEE model, two helices at the both N- and C-terminals look similar to the ones of nns model1 while two helices in the middle look similar to the ones of QUARK model1.



(c) T0827-D2 Initial



(d) T0827-D2 Final



(e) T0827-D2 Alignment

Three topics to cover

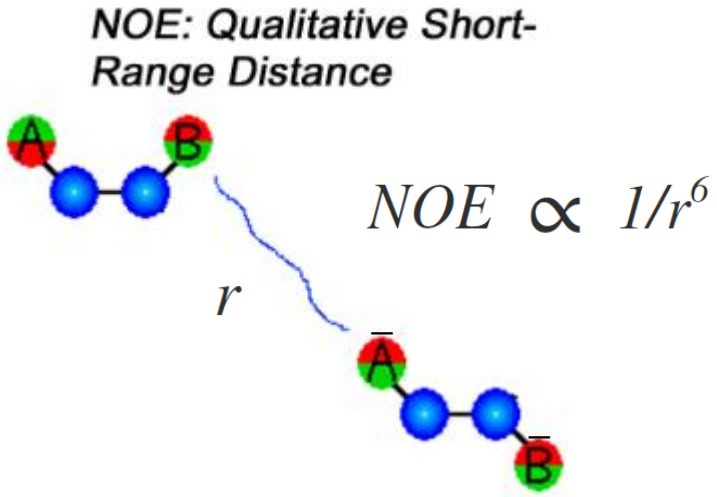
(1) Community detection of a network by modularity optimization

(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

NOE restraints



NOE: "Nuclear Overhauser Effect"

Ambiguous Distance Correction

Ambiguous (trivial & non-trivial) NOEs are those for which more than one assignment is possible.

$$R=\left(\sum_{ij} d_{ij}^{-6}\right)^{-1/6}$$

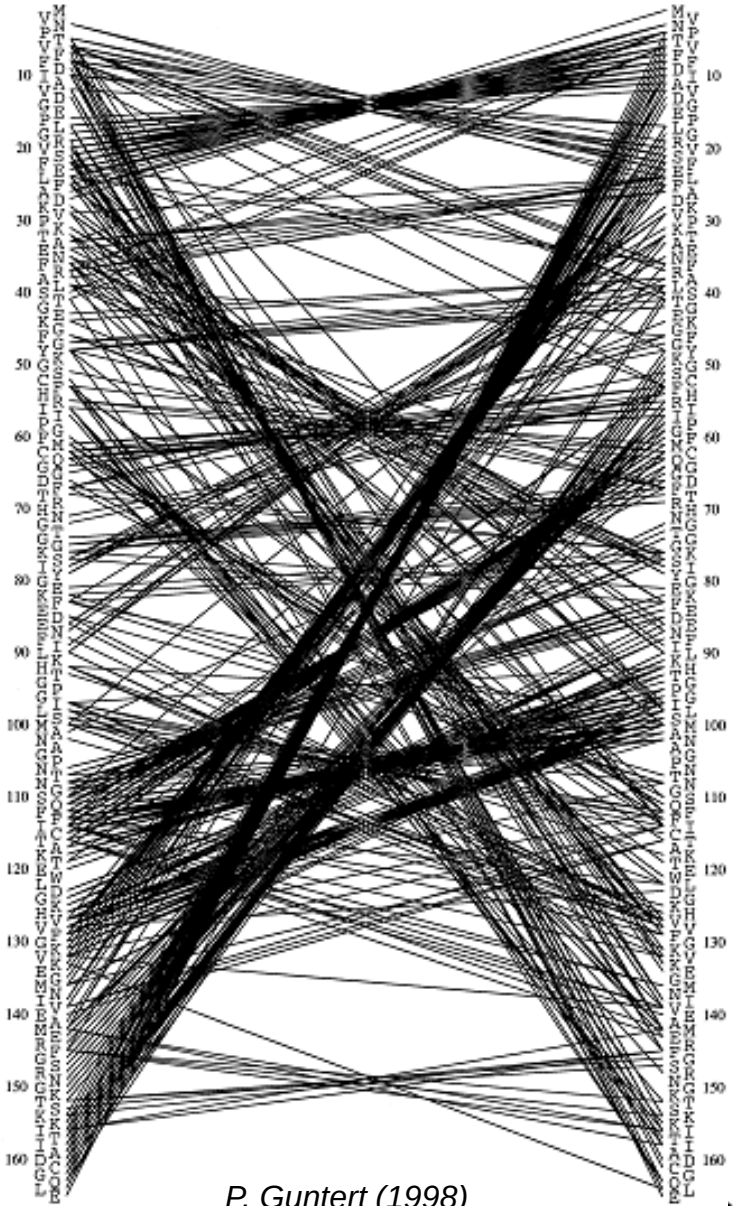
the effective or summed "distance" between more than two atoms.

cf.

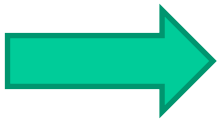
$$R=min\{d_{ij}\}$$

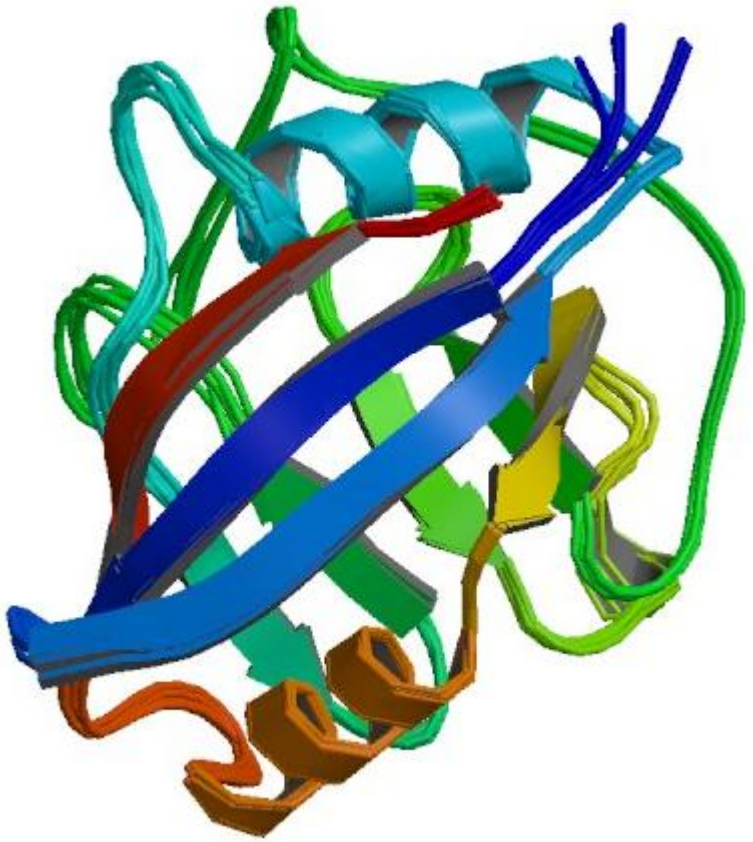
the minimum "distance" between more than two atoms.

5. NMR structure “DETERMINATION”



P. Guntert (1998)





P3. Modeling a protein structure based on NMR data

- Go to BMRB 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 function 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?

Energy Function

$$E_{tot} = E_{NMR} + E_{protein-chemistry} \tag{1}$$

$$E_{NMR} = w_{NOE}E_{NOE} + w_{DIH}E_{DIH} \tag{2}$$

$$E_{protein-chemistry} = E_{stereo-chemistry} + E_{CMAP} + E_{chiral} + E_{repulsive}, \tag{3}$$

$E_{protein-chemistry}$ contains energy terms dictated by the protein chemistry such as $E_{stereo-chemistry}$ to maintain proper bond lengths, bond angles, torsion angles and improper torsion angles, E_{CMAP} [22] to provide a cross-term correction for two adjacent torsion angles in $E_{stereo-chemistry}$, E_{chiral} to keep the chirality of the amino acid residue in the L form, and $E_{repulsive}$ to avoid atomic clashes.

$$E_{repul} = \sum \begin{cases} \epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - 2\left(\frac{\sigma}{r}\right)^{6}\right] + \epsilon & r < \sigma \\ 0 & r \geq \sigma \end{cases}$$

Protein NMR structure "re-determination" by global optimization (under review)

Table II. Comparison of structure qualities between PDB and CSA structure models for 30 targets.

PDB	Nres	NOE (Å)	DIH (°)	Favoured (%)	Outlier (%)	Clash
2G1E	90	0.077/**0.003**	0.460/**0.022**	89.55/ **96.08**	2.05/**0.00**	3.13/**1.00**
2KL7	71	0.006/**0.000**	0.980/**0.019**	90.22/ **96.09**	1.96/**0.65**	12.92/**1.09**
2KO3	76	0.003/**0.001**	0.130/**0.002**	95.41/ **99.39**	0.41/**0.00**	12.99/**0.00**
2RNG	79	0.017/**0.001**	0.190/**0.011**	81.97/ **97.40**	2.13/**1.30**	108.08/**0.00**
2ROG	66	0.169/**0.005**	0.410/**0.007**	92.73/ **97.34**	1.88/**1.33**	0.94/**0.00**
2ROT	70	0.006/**0.001**	-	94.27/ **96.40**	0.22/**0.07**	0.26/**0.00**
2VRD	61	0.020/**0.001**	-	87.78/ **94.41**	0.90/**0.04**	17.44/**0.00**
2YS0	56	0.011/**0.001**	0.060/**0.000**	75.37/ **96.57**	3.06/**0.37**	20.78/**0.00**
2YTV	79	**0.001/0.001**	0.010/**0.006**	92.73/ **96.23**	0.46/**0.00**	0.00/ 0.00
2YUL	82	0.006/**0.002**	0.030/**0.001**	85.44/ **97.25**	0.88/**0.00**	35.94/**1.66**
2YUN	79	0.002/**0.001**	0.020/**0.016**	86.10/ **96.23**	1.10/**0.00**	19.27/**0.00**
2YUO	78	**0.001/0.001**	0.010/**0.008**	90.26/ **97.43**	0.99/**0.20**	13.09/**0.00**
2YUM	75	0.002/**0.001**	0.010/**0.000**	81.58/ **96.99**	1.58/**0.07**	26.78/**0.00**
2PJF	68	0.043/**0.002**	-	84.62/ **93.48**	3.94/**0.00**	102.77/**0.36**
2ROE	66	0.015/**0.001**	-	97.34/ **98.44**	**0.00/0.00**	0.89/**0.00**
2JZ2	66	0.042/**0.001**	-	89.69/ **95.70**	1.72/**0.00**	16.14/**0.28**
1XJH	62	0.020/**0.001**	-	93.33/ **98.33**	**0.00/0.00**	0.32/**0.00**
2YUZ	100	0.042/**0.001**	0.050/**0.027**	88.47/ **98.83**	1.22/**0.15**	15.44/**0.00**
2YR3	99	0.002/**0.001**	0.080/**0.003**	90.21/ **95.93**	1.24/**0.05**	11.72/**0.00**
2NOC	99	0.024/**0.001**	-	85.77/ **95.77**	2.78/**0.00**	18.55/**0.32**
2YU0	94	0.002/**0.001**	0.020/**0.003**	84.19/ **97.34**	0.44/**0.00**	23.12/**0.78**
2YUP	90	**0.001/0.001**	0.010/**0.003**	85.40/ **98.86**	1.25/**0.00**	16.68/**0.00**
2YUK	90	0.003/**0.002**	0.010/**0.000**	88.13/ **98.92**	0.57/**0.00**	32.01/**1.15**
2KO6	89	0.018/**0.001**	0.550/**0.007**	90.52/ **98.91**	1.49/**0.00**	14.65/**0.00**
2KL1	87	0.062/**0.002**	0.740/**0.008**	90.71/ **99.77**	2.71/**0.00**	13.63/**0.00**
2YUQ	85	**0.001/0.001**	**0.000/0.000**	89.52/ **96.08**	**0.60**/1.21	11.18/**0.00**
2KL8	85	0.004/**0.000**	0.300/**0.006**	94.70/ **98.13**	0.48/**0.00**	13.67/**0.25**
2ROW	84	0.004/**0.001**	0.230/**0.007**	78.66/ **95.97**	5.49/**1.22**	27.07/**2.90**
2NRG	82	0.022/**0.004**	0.860/**0.031**	77.50/ **97.50**	6.25/**1.25**	50.15/**0.77**
2YQI	81	0.003/**0.001**	0.205/**0.002**	91.84/ **99.94**	0.70/**0.00**	21.58/**0.00**
Ave		0.021/**0.001**	0.233/**0.008**	88.13/ **97.19**	1.62/**0.26**	22.04/**0.35**
Std		0.033/**0.001**	0.299/**0.009**	5.45/ **1.55**	1.52/**0.47**	25.48/**0.66**

Ts targets of CASP11

- Sparse NMR distance restraints that reflect data available in the initial stages of the state-of-the-art NMR study of a large protein is provided.
- Many restraints are ambiguous. For each NOESY peak one or more distance restraints are provided, of which at least one is correct.
- The corresponding constraints are sparse and usually not sufficient to refine the structure using standard NMR packages.
- The challenge for us is to either solve the structure using more sophisticated modeling techniques or to provide at least partially correct models, facilitating interpreting more complex NMR data sets.

19 Ts targets in CASP11

Targets	Nres	Npeaks	Npeaks/residue	Max Npair	Avg Npair	Avg upper (A)
Ts761	237	3106	13	540	36	7.9
Ts763	130	2029	15	270	22	8.0
Ts777	345	2400	6	1296	71	5.9
Ts785	112	694	6	351	29	6.2
Ts794	462	3132	6	2232	122	6.1
Ts800	212	1459	6	1053	74	6.1
Ts802	118	530	4	135	14	6.0
Ts804	194	884	4	1395	43	5.9
Ts810	113	739	6	270	18	5.9
Ts814	397	2290	5	1314	69	6.0
Ts818	134	516	3	162	13	5.6
Ts824	108	522	4	207	9	5.8
Ts767	274	1564	5	396	34	5.9
Ts806	256	1791	6	1368	88	6.1
Ts812	183	980	5	684	29	6.1
Ts826	201	1666	8	2448	96	6.0
Ts827	158	1091	6	918	46	6.0
Ts832	209	1472	7	1035	68	6.0
Ts835	404	3517	8	2106	106	6.1
Average	224	1599	6	957	52	6.2

* Npeaks/residue ~ 15, Avg Npair ~ 2 (for 30 PDB)

Tsc protocol: two-level optimization problem

$$E = E_{NOE} + E_{stereo\text{-}chemistry} + E_{vdw}^{repul} + E_{chiral} + E_{CMAP}$$

$$E_{NOE} = \sum_{noe} \begin{cases} k(R-d_{max})^2, & R>d_{max} \\ k(R-d_{min})^2, & R<d_{min} \\ 0, & in\ \ between \end{cases}$$

Chiral torsion (CA-N-C-CB) ~ +35

$$E_{chiral} = \sum_{chiral} (\chi-\chi_0)^2$$

Two-adjacent-dihedral-angles cross-term (CMAP from CHARMM22)

$$E_{CMAP} = \sum_{\phi,\psi} u(\phi,\psi)$$

***Initial structures:** 1) fold recognition models generated by the early stage of nns. 2) other server models after clustering.

***2-level optimization:** 1) NOE assignment and 2) 3D model generation
***Single re-minimization:** LBFGS using E' = E + $E_{electro\text{-}static}$ + E_{gbsa}
***Model selection:** 1) by clustering 2) higher ramachandran favored score, lower outlier, lower clash score (lower restraint violation)

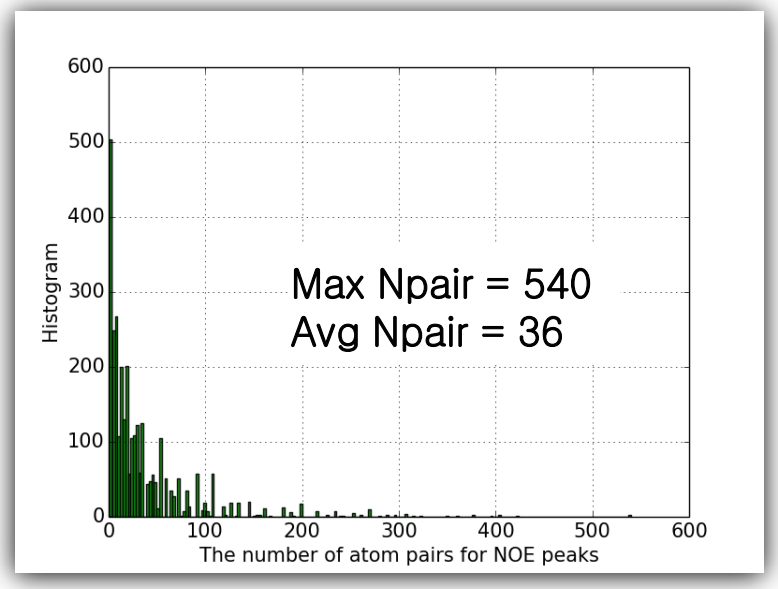
The first target… Ts761

MMKLARKSVPFIIAVALLAACLLAVGLSPLVLPDYKGTIEEREQPQNFNLLYLNSGEELNLYPWNLYTGQEQELFEEEIVSFAANSVRI
LGGGSWTDEELYPLIKFRYSGQDLRFLKDMALTEKDGRRYLVNMALDPNGLCYFSYVNQDEREATADEMDQALGKLQEDWEKFLSDPLP
ADSEVDLYEEKPSGSYQLDDGELKTDNAFYMFFMRCQMLSDQMRKEQYSDYIGDNLYTIWELVLKSEFTSLSYDNHIYAMYSNDGGTSM
VLIYSPIEERFVGFSLKY

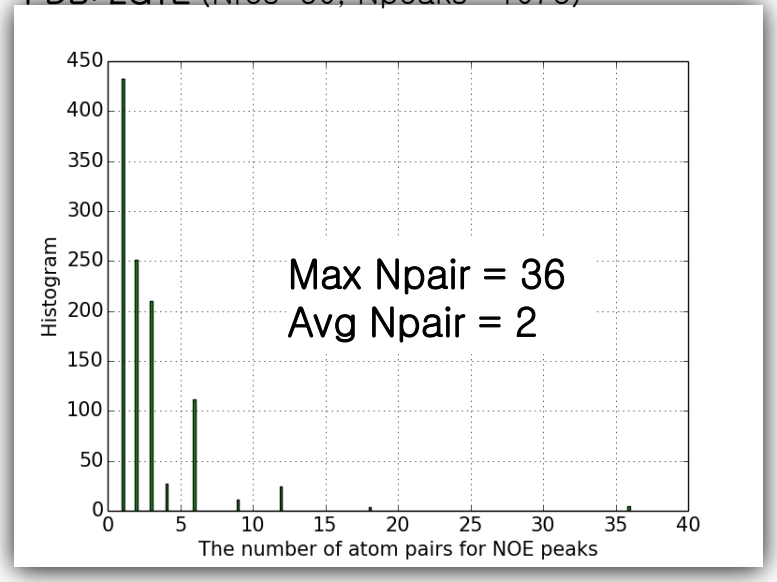
Nres = 237 aa

Npeaks = 3106 (= the number of NOE distance restraints)

Ambiguity of NOE peaks



PDB: 2G1E (Nres=90, Npeaks= 1076)

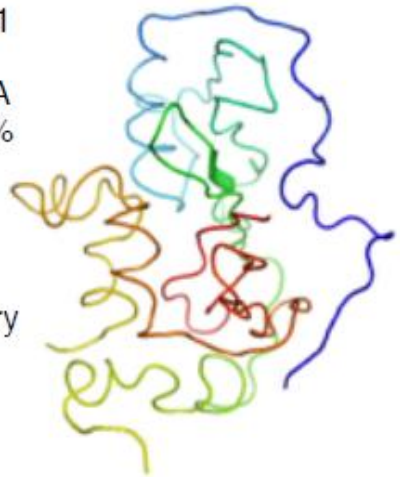


Ts761 (237 aa)

Model1

NOE ~ 0.002 A
Favored ~ 62%
Outlier ~ 11%

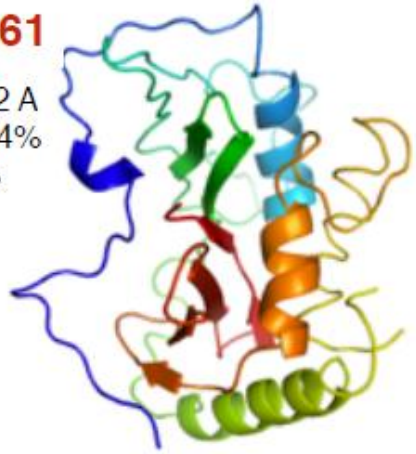
poor secondary structures



RMSD ~ 17 A

Tc761

NOE ~ 0.002 A
Favored ~ 84%
Outlier ~ 5%



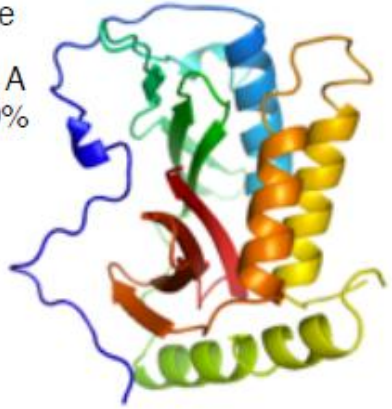
Re-calculated

NOE ~ 0.000 A
Favored ~ 94%
Outlier ~ 1%

RMSD ~ 2.1 A

native

NOE ~ 0.000 A
Favored ~ 99%
Outlier ~ 0%

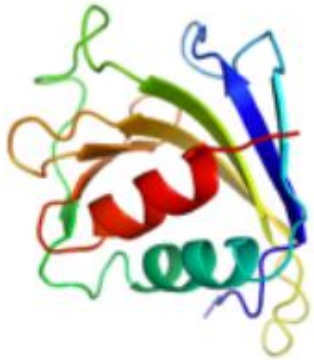


Results for 19 Ts targets (LEE)

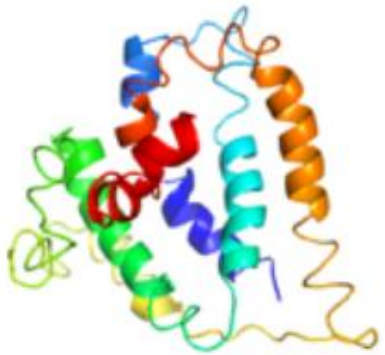
*TM-score & RMSDs: **Ts**
Reds: using native structure
otherwise: using Tc structures

Targets	Nres	TM-score	RMSD (A)	NOE (A)	Favoured (%)	Outlier (%)	Clash	
Ts761	237	0.9084	2.11	0.000	93.81	0.95	0.00	recalculated
Ts763	130	0.8589	2.15	0.001	89.84	1.56	0.00	
Ts777	345	0.7113	5.60	0.002	77.26	9.04	0.00	recalculated
Ts785	112	0.8290	4.00	0.001	94.55	1.82	0.00	
Ts794	462	0.8290	4.35	0.002	86.52	3.26	0.28	
Ts800	212	0.9334	1.60	0.000	93.33	0.48	0.00	
Ts802	118	0.7815	2.68	0.000	89.66	3.45	0.00	
Ts804	194	0.7350	4.90	0.000	89.06	4.17	0.00	
Ts810	113	0.7478	5.18	0.000	92.79	0.00	0.00	
Ts814	397	0.9125	2.57	0.001	86.84	5.32	0.00	
Ts818	134	0.8676	2.05	0.001	92.42	0.76	0.00	
Ts824	108	0.8332	2.32	0.001	94.34	0.94	0.00	
Ts767	274	0.8958	2.33	0.002	87.87	1.47	0.45	
Ts806	256	0.8905	2.39	0.001	90.55	1.58	0.00	
Ts812	183	0.8003	3.40	0.001	82.32	4.97	0.00	
Ts826	201	0.7722	4.54	0.002	87.44	1.01	0.00	
Ts827	158	0.4627	8.90	0.003	80.77	3.21	0.00	2 domain (orientation is different)
Ts832	209	0.8410	2.53	0.002	87.92	2.42	0.00	
Ts835	404	0.9398	1.93	0.004	90.30	2.74	0.62	
Average	224	0.8184	3.45	0.001	88.82	2.59	0.07	

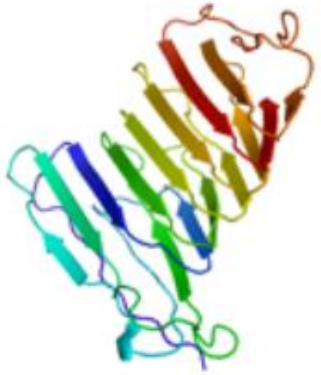
Structure Examples



Ts763 (130aa) ~ 2.1 A

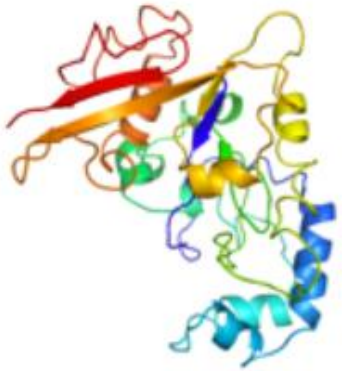


Ts826 (201aa) ~ 2.1 A



Ts800 (212aa) ~ 1.6 A

Ts767 (274aa) ~ 2.3 A



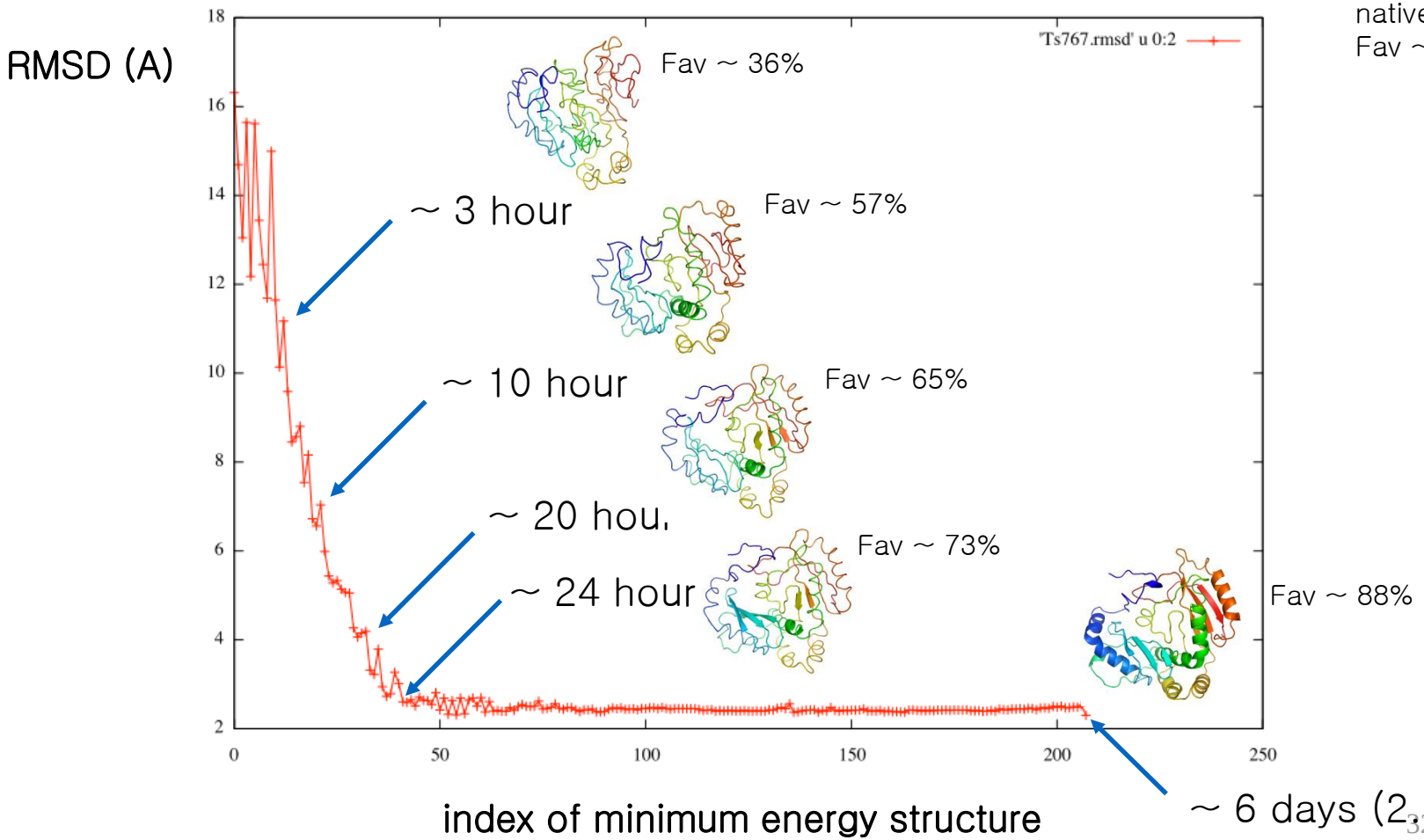
Ts806 (206aa) ~ 2.3 A

Ts835 (404aa) ~ 1.8 A

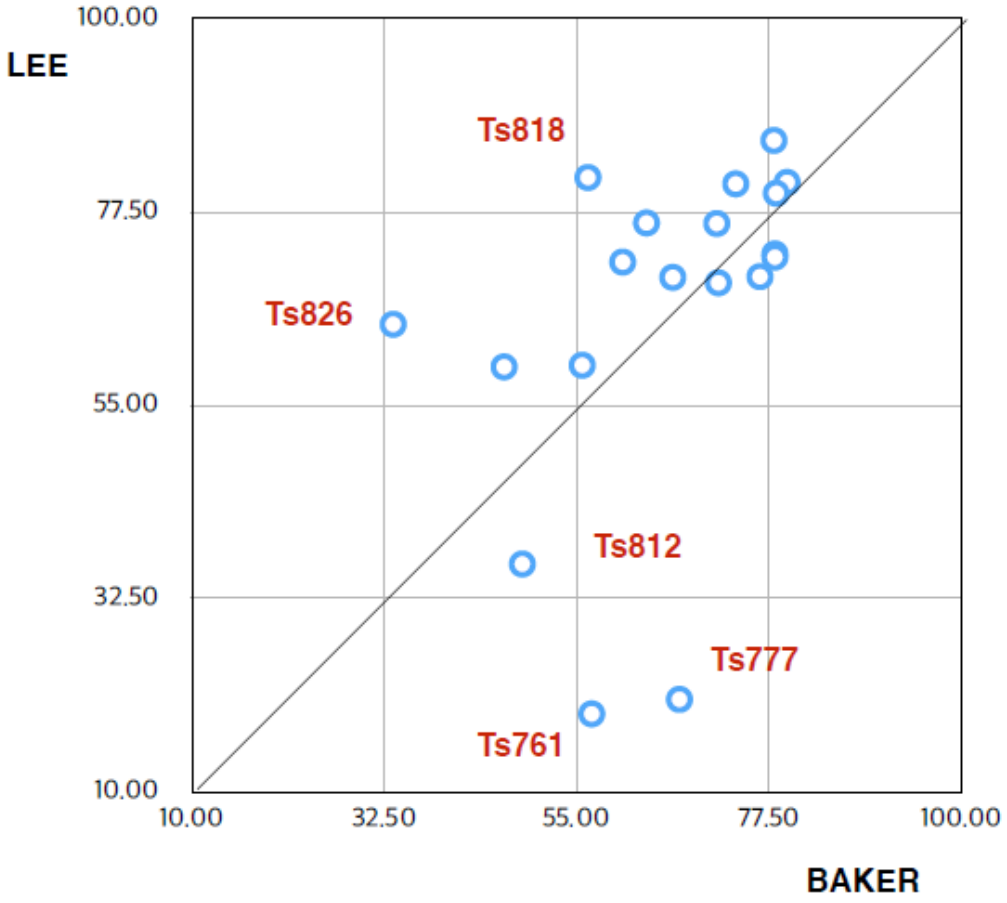
Modeling of Ts767 (274 aa)

native
Fav ~ 98%

~ 200 CPU cores of Intel Xeon X5670 at 2.9 Ghz



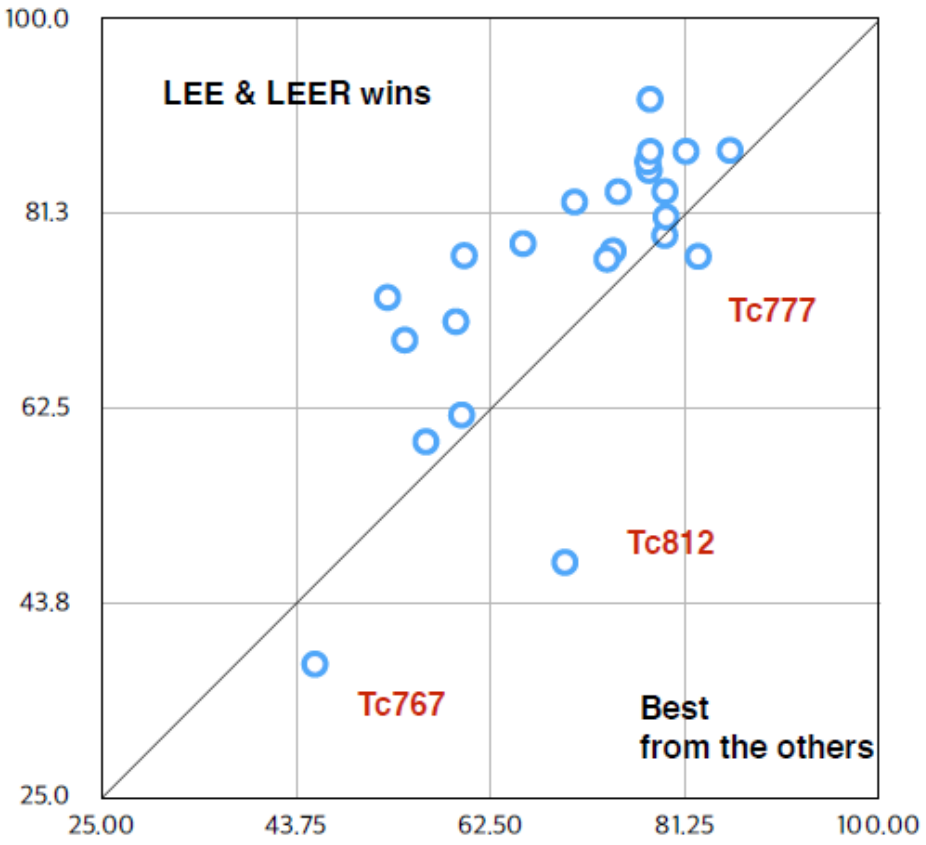
LEE vs BAKER

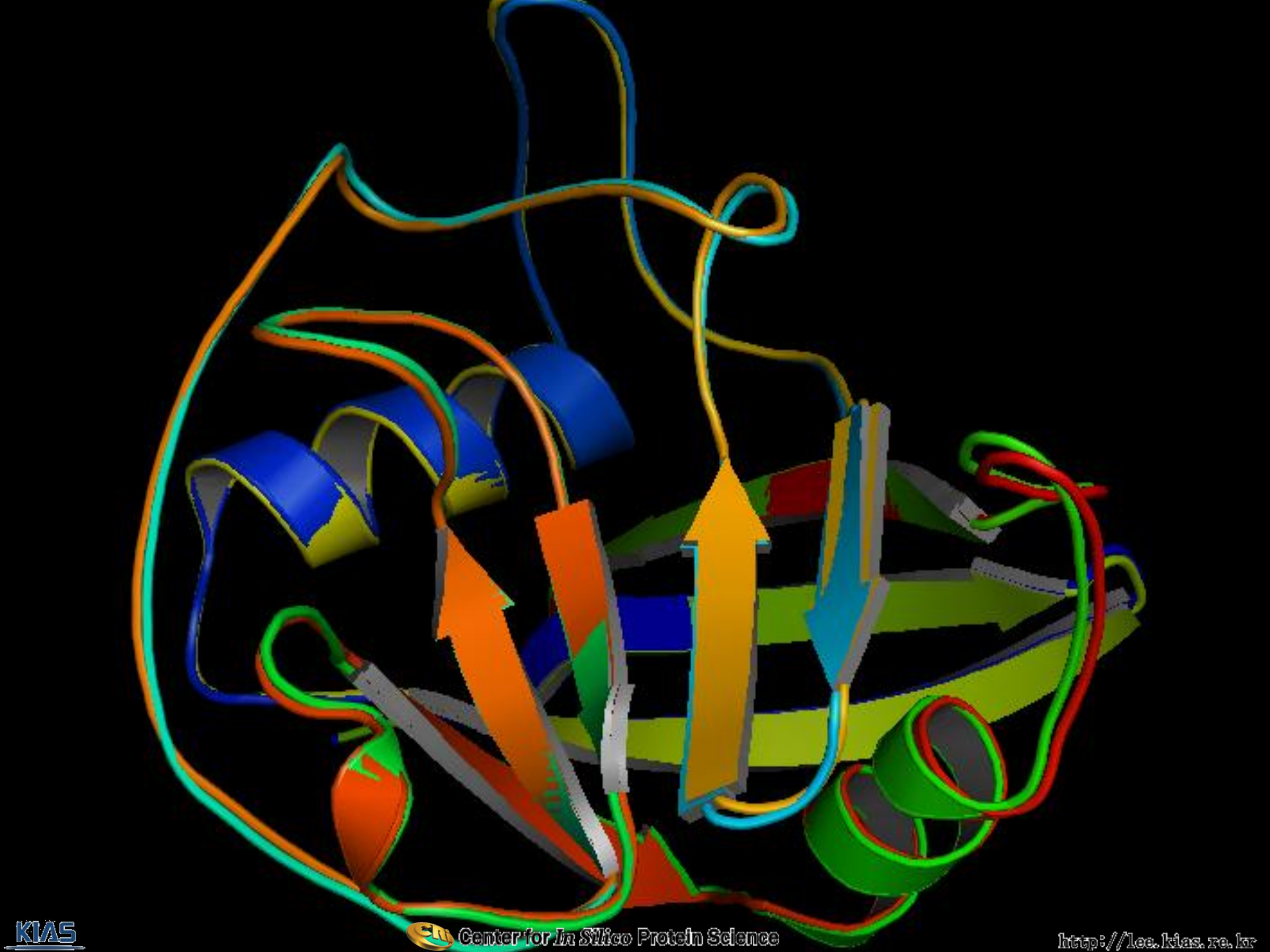


*19 Ts targets
*Official GDT-TS score

*Ts761 & Ts777:
Failure is due to dum "smart" atom-pair screening

LEE(R) vs best of the others (20:4)

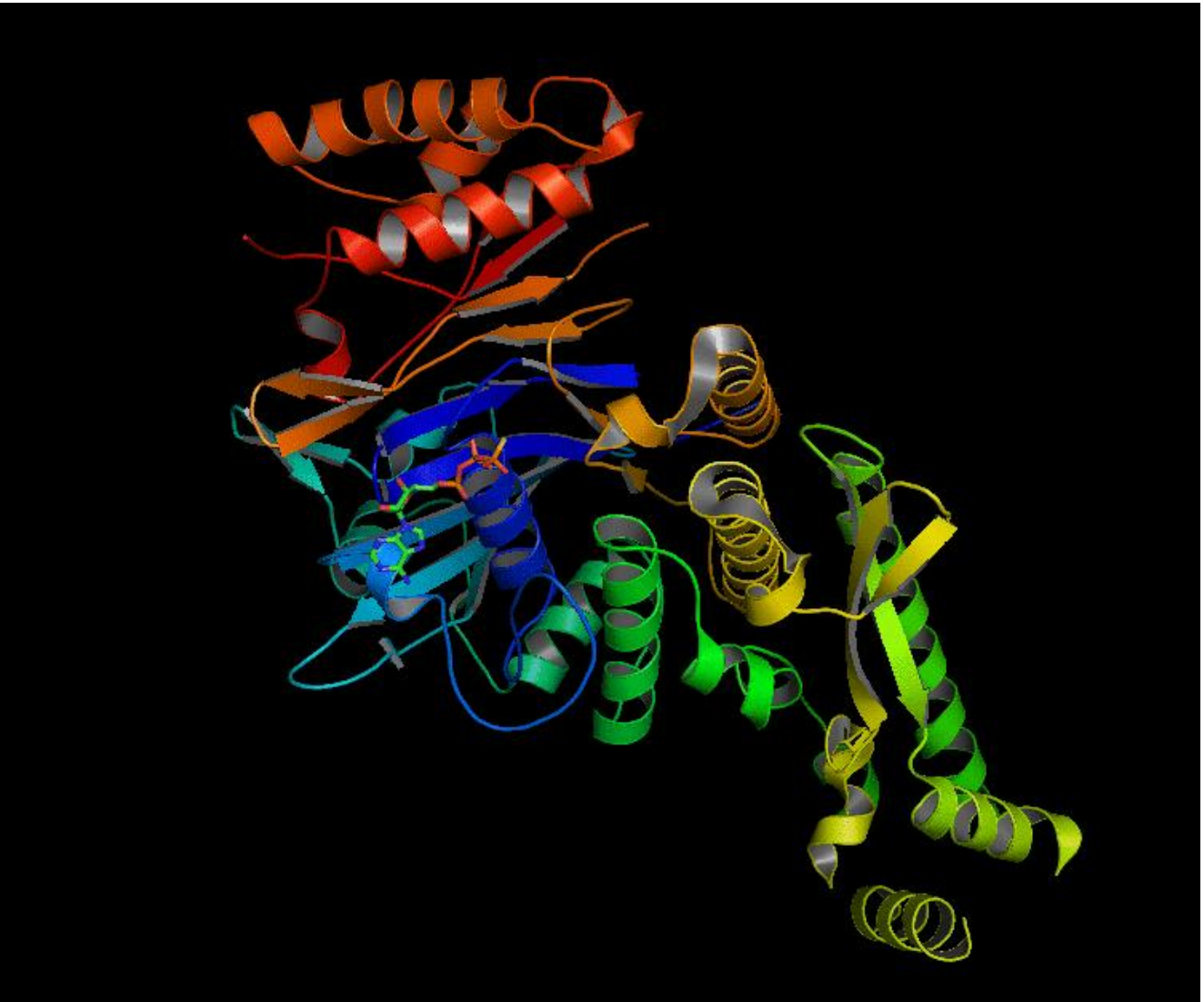



KIAS
Center for In Silico Protein Science
http://lee.kias.re.kr

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)



Protein folding problems

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

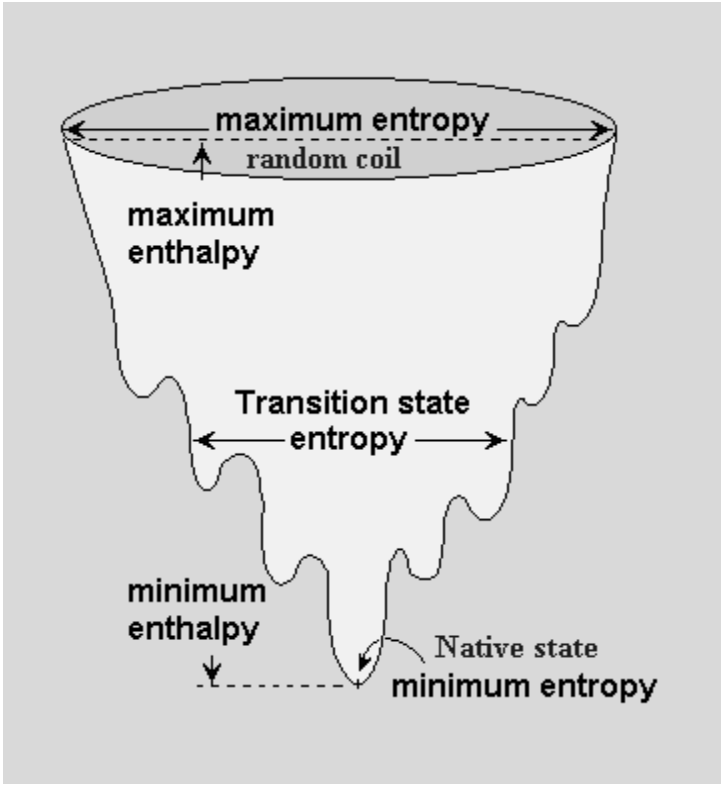
Protein-Folding Mechanisms

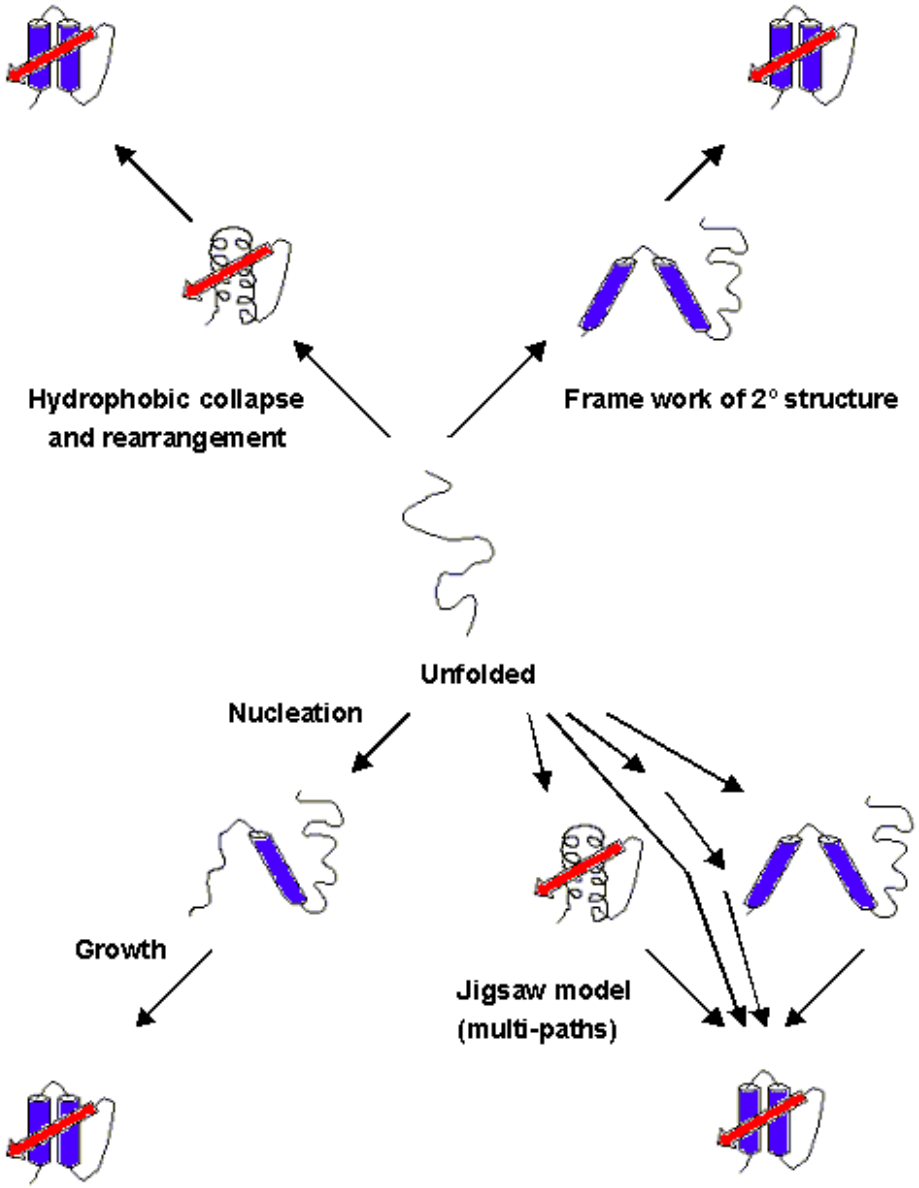
Random search of all conformational space requires an immense amount of time (longer than the age of universe). ←→ In vitro refolding normally takes seconds or minutes.

–Levinthal paradox

1. Consider a small protein with 100 amino acids.
2. If we assume that each residue can take a structure out of three secondary element of helix, sheet, and coil, the total number of possible structures of this protein is 3^{100} (or about 10^{48}).
3. The time scale for a residue to reshape from one SS to another $>10^{-14}$ sec (time resolution for a bond to rotate)
4. Time to find the native structure of the protein by random search $>10^{34}$ sec (10^{26} year) → Longer than the age of universe!
5. Conclusion: There must be folding pathways!!!

→ Folding pathway problem: identifying intermediates and constructing folding mechanism





Various models for protein folding

Computational Studies on Protein Folding Mechanisms

→**No success yet on simulating protein-folding processes.**

Existing approaches:

(1) direct folding simulations (e.g. Kollman's 1μs MD on HP-36):

no foldings yet observed in simulation: accuracy of the potential energy?

recent MD simulations using Anton (mili-second simulations): Science v334 page 517 & Biophys J. v100 page L47.

(2) simple (lattice, minimal, …) models: HP, BLN, etc

trying to understand principles of protein folding

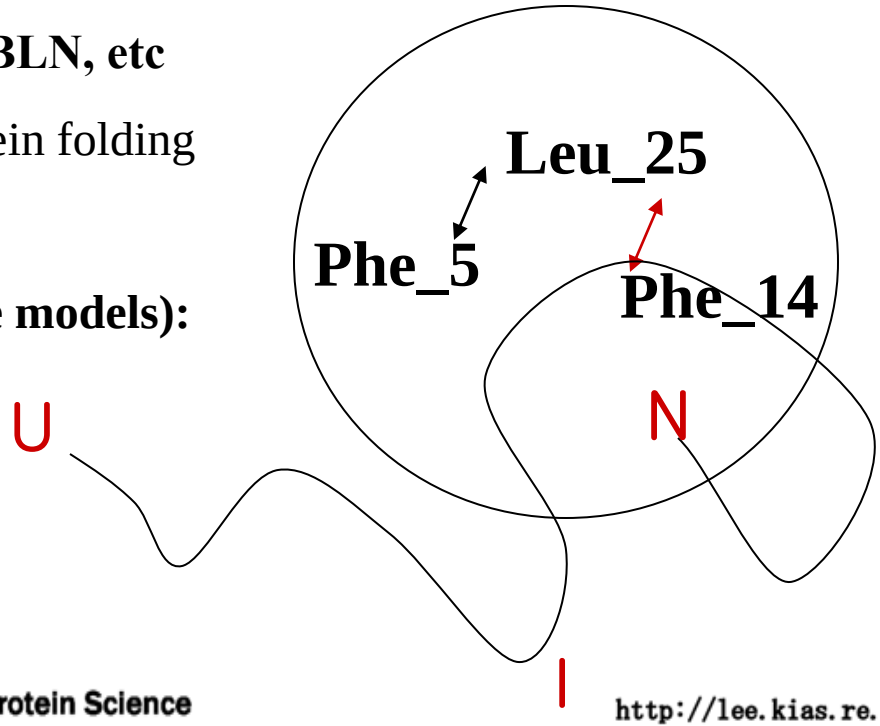
not realistic

(3) consider only native interactions (Go-type models):

not realistic

(4) unfolding simulation:

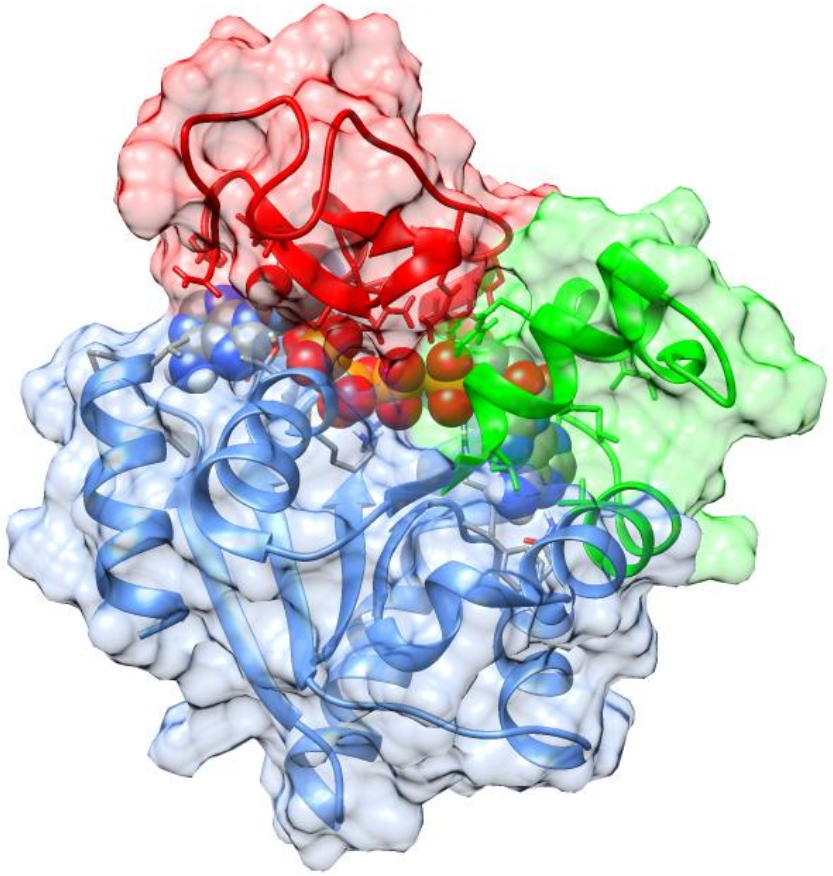
folding is the reverse of unfolding ???



Adenylate Kinase (AdK)

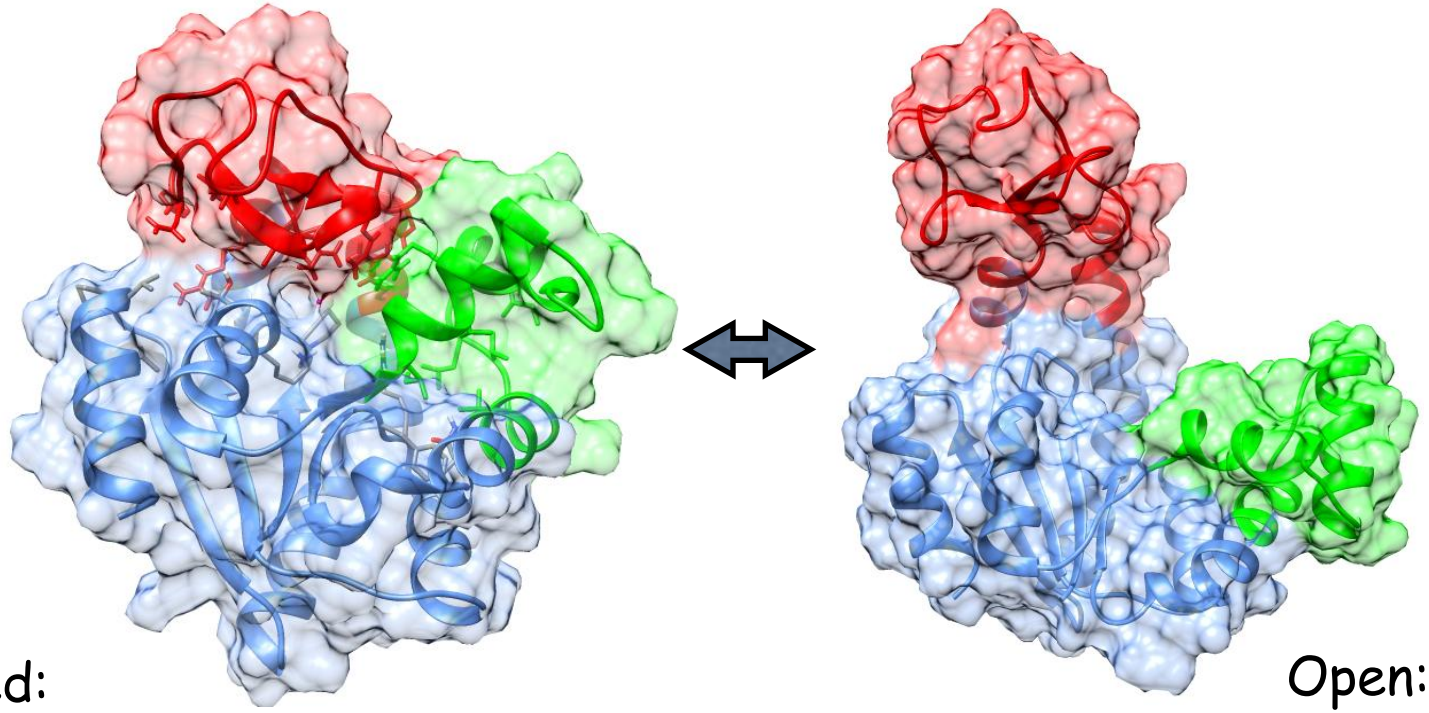
Lee, J.,Joo, K., Brooks BR, Lee, J. (in press)

- One of the most investigated systems for conformational changes
- Phospho-transferase enzyme
- 2ADP ←→ ATP + AMP
- Essential in cellular energy homeostasis
- LID, NMP and CORE domains



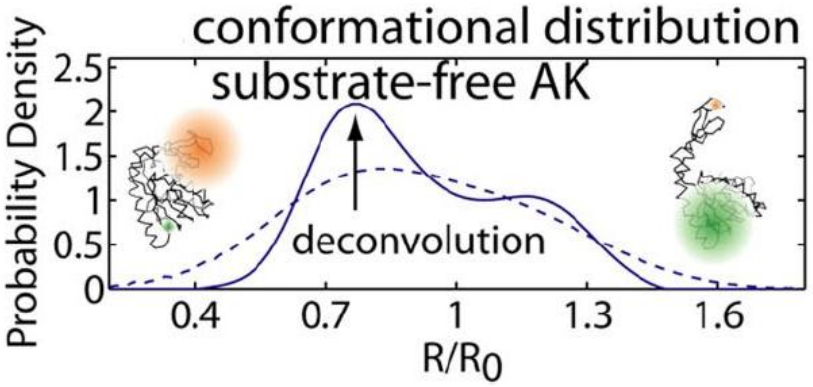
PDB ID: 1AKE

Mechanism



Closed:
1AKE

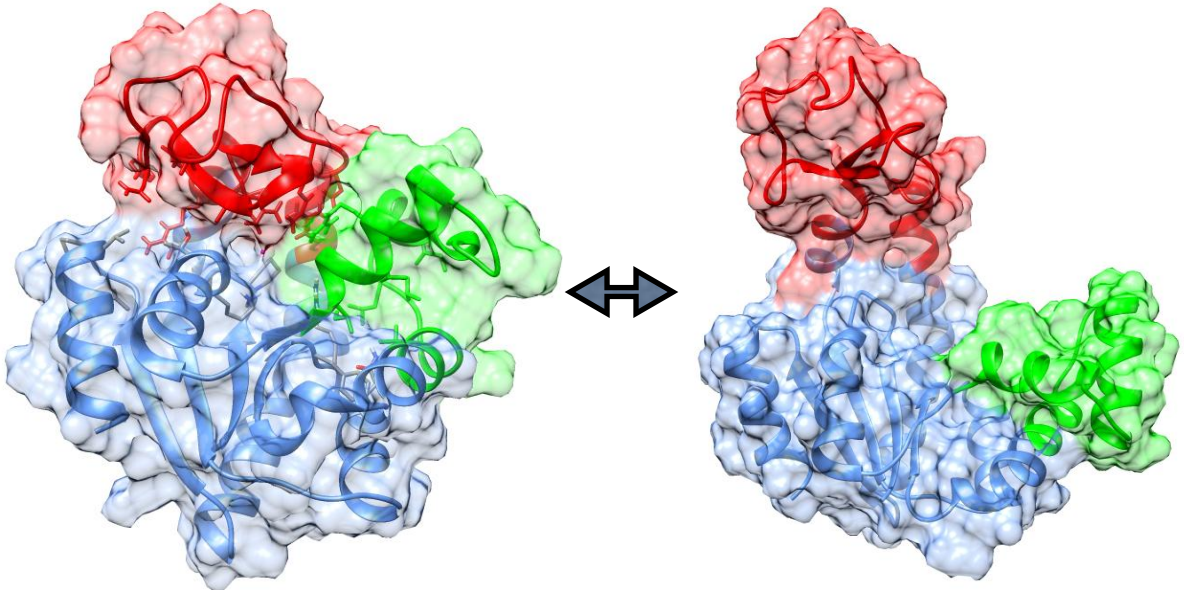
Open:
4AKE



Hansen et al. (2007) *PNAS* **104** 18055

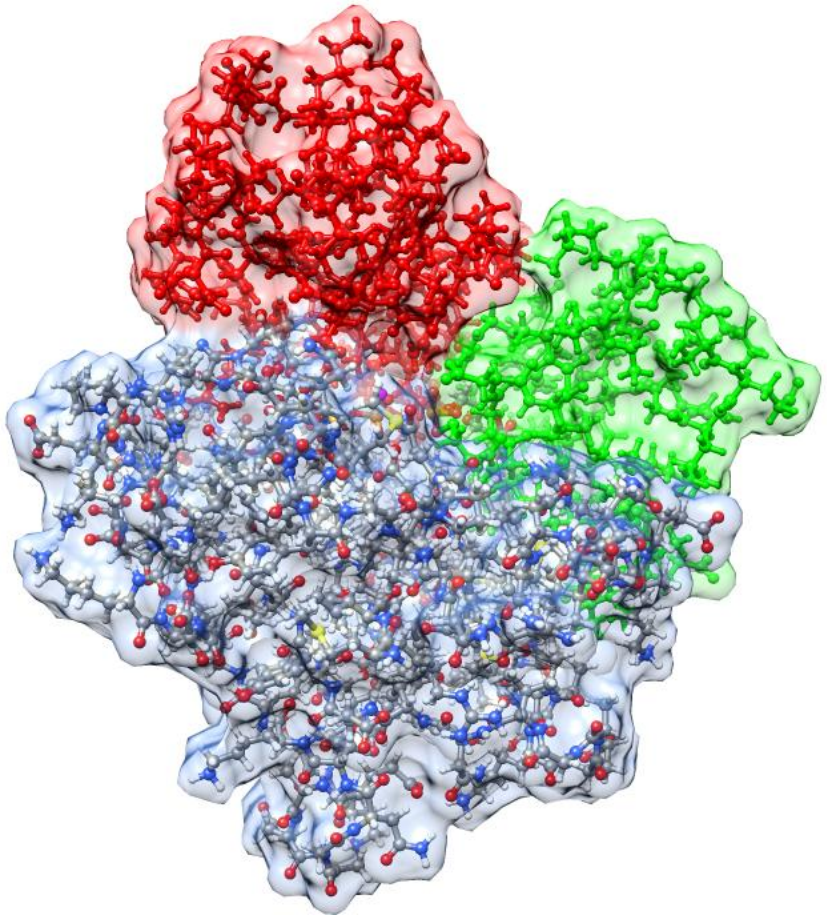
Questions

1. Which residues are crucial for the conformational change?
2. What is the transition state(s)?
   - LID domain first vs. NMP domain first
   - symmetric pathway vs. asymmetric pathway
3. Intermediate state?
4. Comparison with NMR amide-bond fluctuation experiment data



All-atom simulations

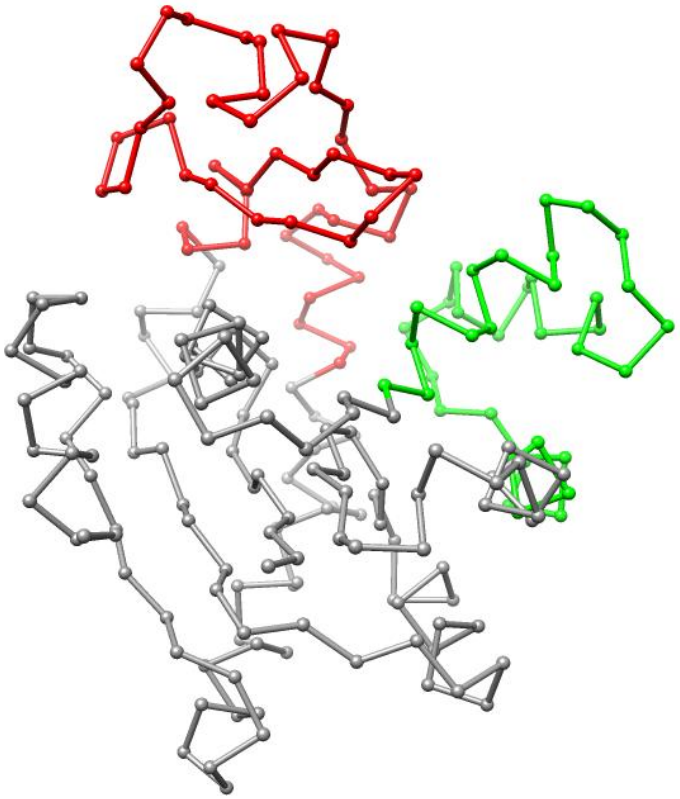
- MD + Principal component analysis
  - RI Cukier, (2006) *JPCB*
- Umbrella sampling
  - Arora & CL Brooks, (2007) *PNAS*
- Dynamics importance sampling
  - Beckstein et al., (2009) *JMB*
- 100 ns MD simulation
  - Brokaw & Chu, (2010) *Biophys.*
- Minimum energy path
  - Matsunaga et al., (2012) *PLoS Comput. Bi*



Only 1~2 transitions are observed!!!

Coarse-grained models

- Protein is represented by C_α trace.
- Pseudo angle/dihedral angle as well as C_α - C_α distance restrains are used.
  - Mixed plastic network model: P. Maragarkis & M. Karplus, (2005) JMB
  - Double-well network model: J. Chu & G Voth, (2007) Biophys.
  - Structure-based model/Gō-model:P. Whitford et al., (2007) JMB
  - Mixed Gō-model: MD Daily et al. (2010) JMB



Chemical details are missing and consequently agreement with experimental data is either rather limited or none.

$$H = \frac{1}{2} k \sum_{i} \sum_{\substack{j>i \\ R_{ij}^{native} < R_c}} (R_{ij} - R_{ij}^{native})$$

GOAL

- To perform straightforward **all-heavy-atom MD** simulations (solving an initial value problem).
- To observe **numerous spontaneous conformational transitions of AdK** between the open state and the closed state for atomistic investigation of conformational changes of AdK.
- Necessary conditions:
  - need to stabilize two given structures (open and closed) → structure-based modeling (Go model)
  - need to establish a minimal free energy barrier between two states → proper mixing of two given structures

Our approach with all-heavy atom representation of protein

$$H = H_{stereochemistry} + w_{vdW} H_{vdW} + w_{contact} H_{contact}$$

$$= \sum_{bonds} K_b (b - b^0)^2 + \sum_{angles} K_\theta (\theta - \theta^0)^2 + \sum_{dihed} K_\phi (1 + \cos(n\phi - \delta))$$

$$+ \sum_{improper} K_\omega (\omega - \omega^0)^2 + w_{vdW} \sum_{\substack{ij \\ i \neq j}} \varepsilon_{ij} \left(\left(\frac{r_{ij}^{\min}}{r_{ij}} \right)^{12} - \left(\frac{r_{ij}^{\min}}{r_{ij}} \right)^{6} \right)$$

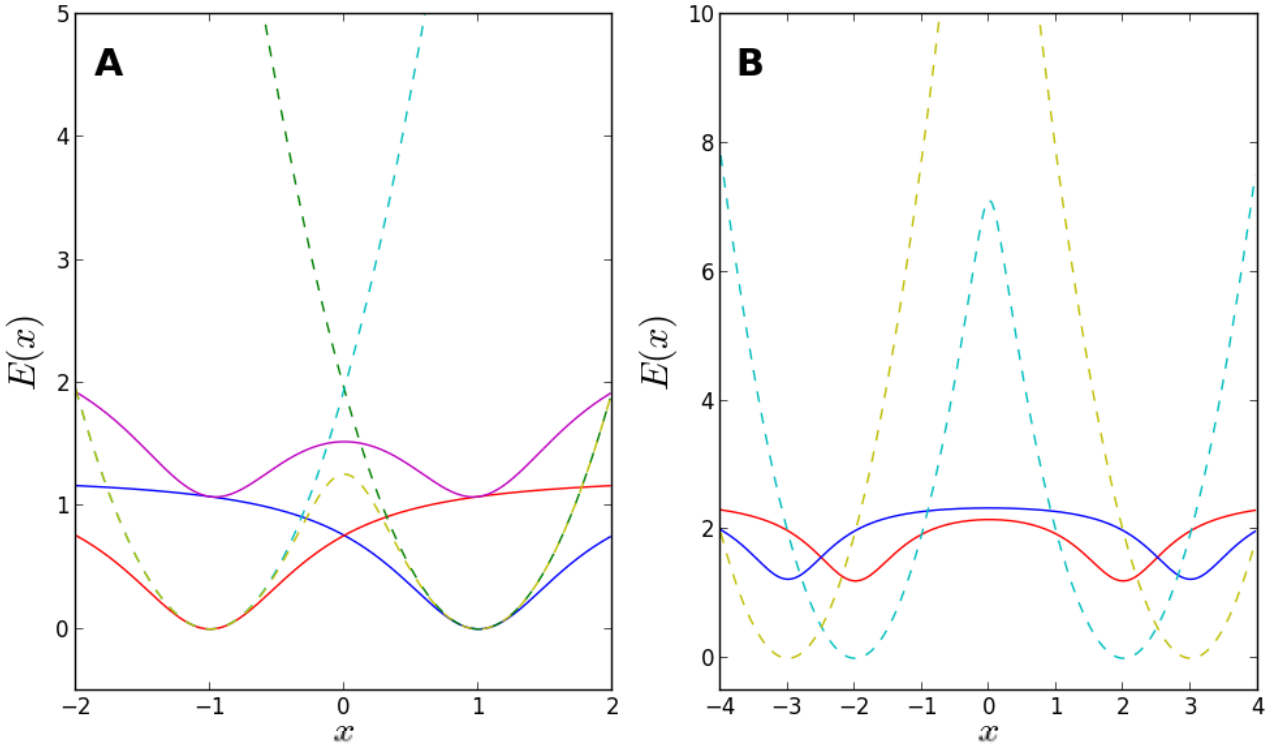
Physics-based terms

$$+ w_{contact} \sum_{contacts} \frac{1}{\sigma_{ij}} \frac{(r_{ij} - r_{ij}^{native})^2}{(r_{ij} - r_{ij}^{native})^2 + \sigma_{ij}^2}$$

Structure-based terms

1. Start with a PDB structure of AdK.
2. T=300K, equilibration MD simulation of 2 ns is performed.
3. A total of 300 straightforward separate MD simulations (T=300K) are performed starting from a randomly perturbed structure. → A total of 6μs MD simulation is performed where RMSD is measured from two native structures of AdK [*BMC Bioinformatics, 16, 94 (2015)*]

Lorentzian structure-based mixing (our approach) vs. coventional Boltzmann weighted mixing scheme



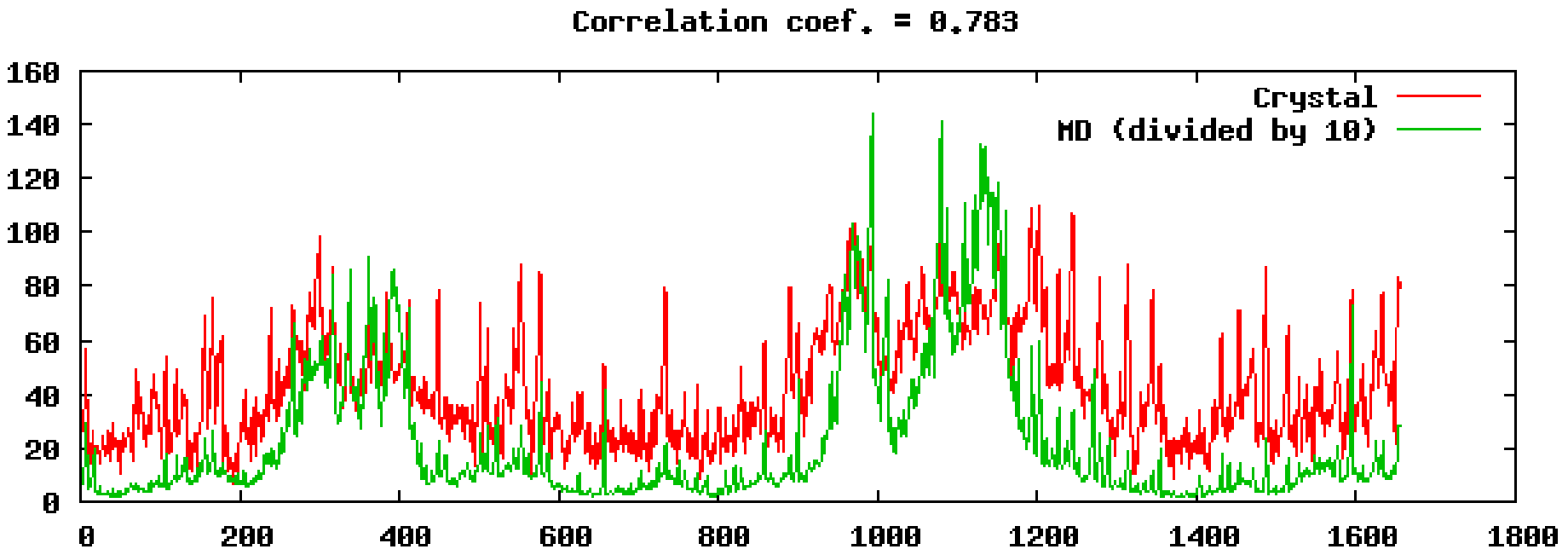
$$L(R_{ij};R_{ij}^{0},\sigma)=\sum\frac{1}{\sigma}\frac{(R_{ij}-R_{ij}^{0})^{2}}{(R_{ij}-R_{ij}^{0})^{2}+\sigma^{2}}$$

$$H(R_{ij})=-\beta\ln[e^{-\beta(R_{ij}-R_{ij}^{'})^{2}}+e^{-\beta(R_{ij}-R_{ij}^{''})^{2}}]$$

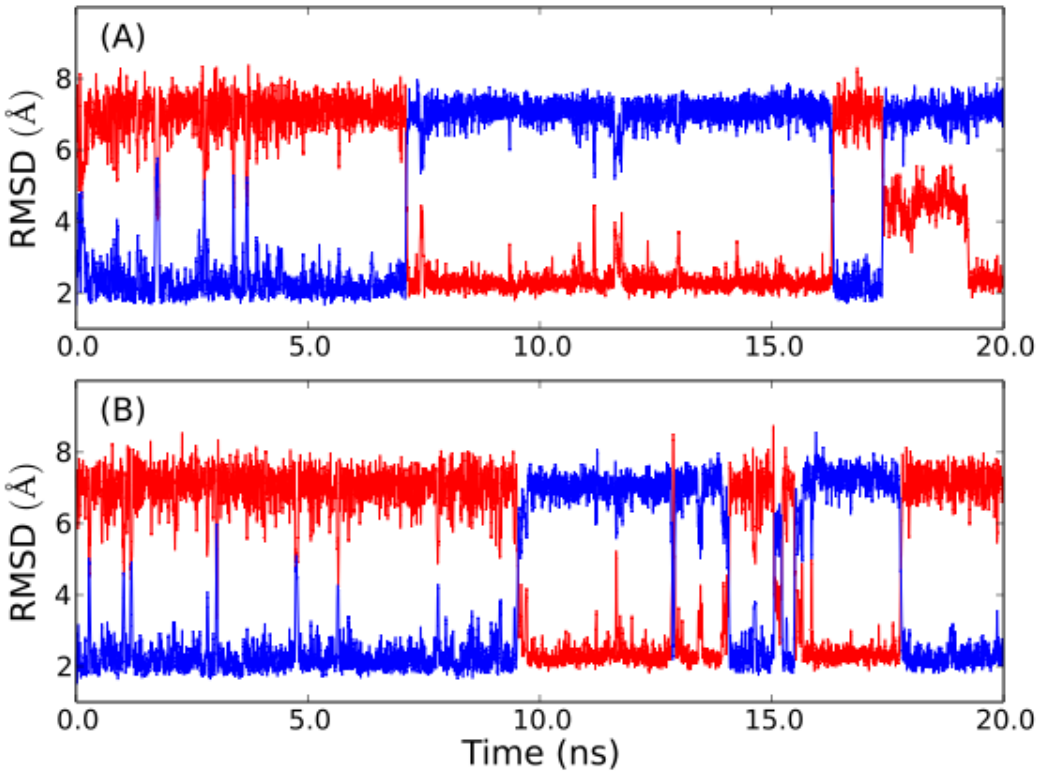
Best & Hummer, *Structure* (2005)
Daily, Phillips & Cui, *JMB* (2010)

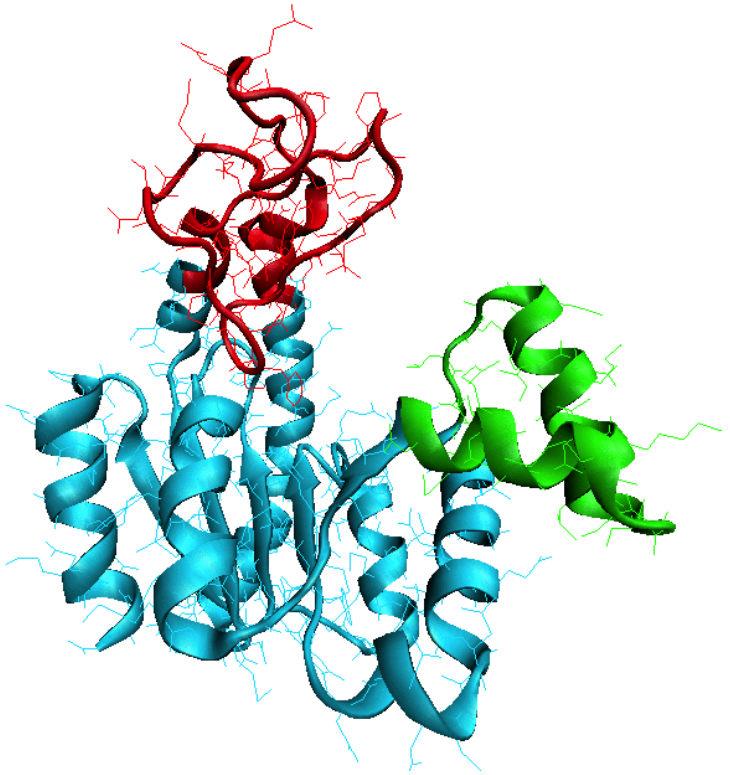
Benchmark I:

Comparison of atomic B-factor values between experiment and simulation shows improved results in correlation than existing methods.



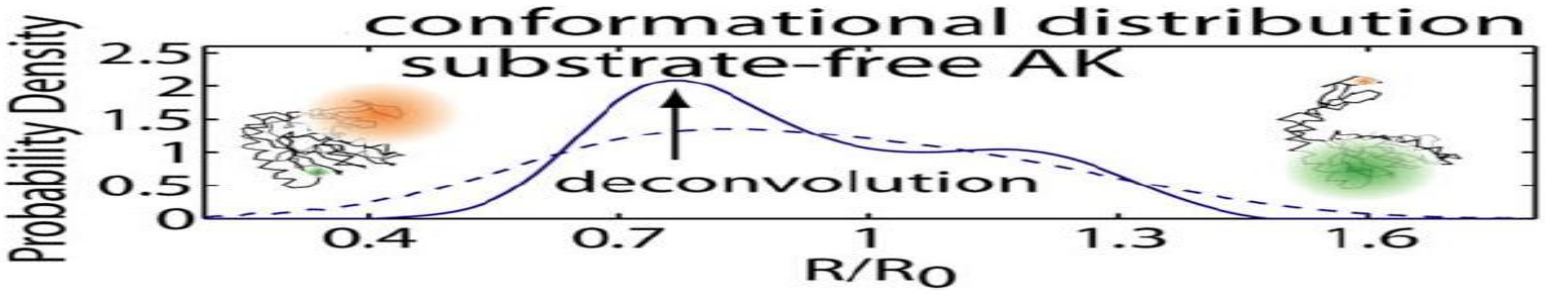
Multiple Spontaneous Transitions

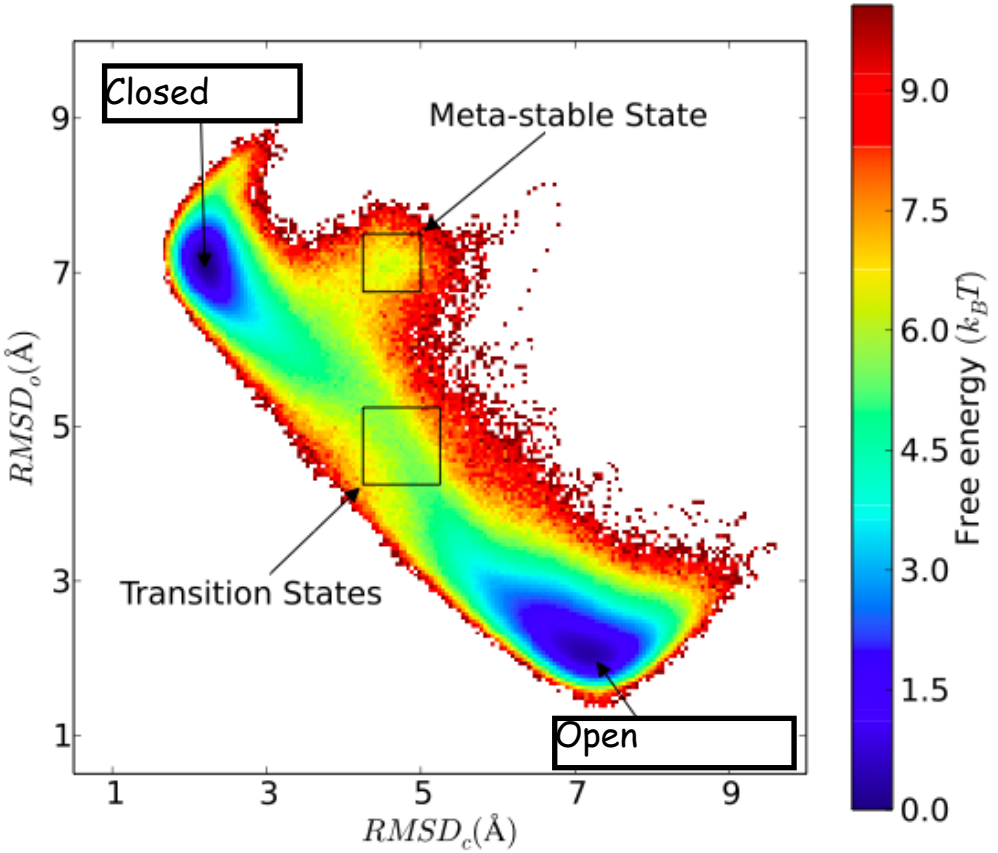




Over 1,000 spontaneous transitions are observed during 6μs MD simulation

Free energy landscape



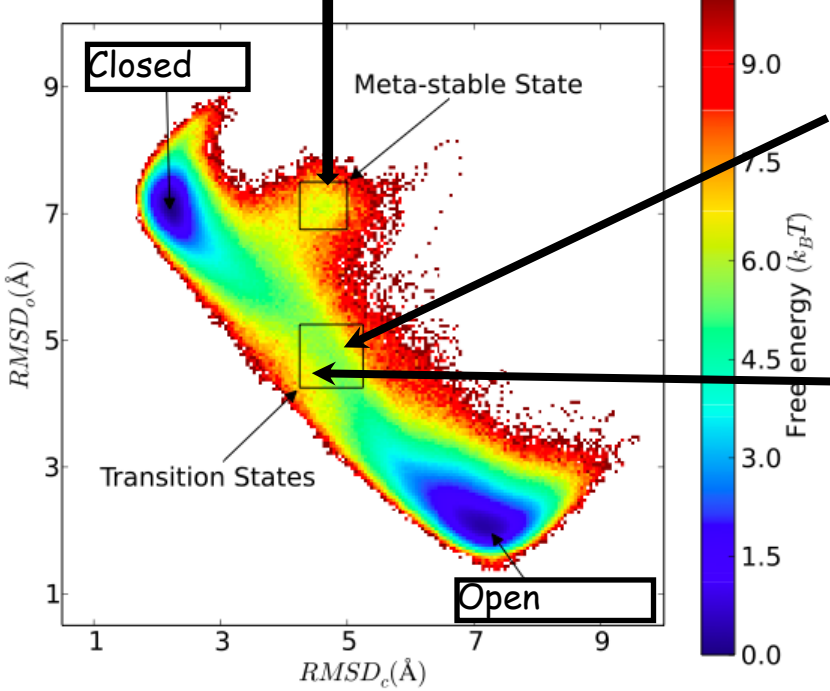


Structural clustering analysis of TS and MS

Conformations in TS and MS are separatly used to construct two non-weighted networks using 2Å RMSD cutoff:

- TS network is divided into 2 giant-component disconnected clusters
- MS → 1 giant-component cluster

LID-twisted
NMP-closed



LID-closed
NMP-open

LID-open
NMP-closed

LID-closed & NMP-open conformation is dominant

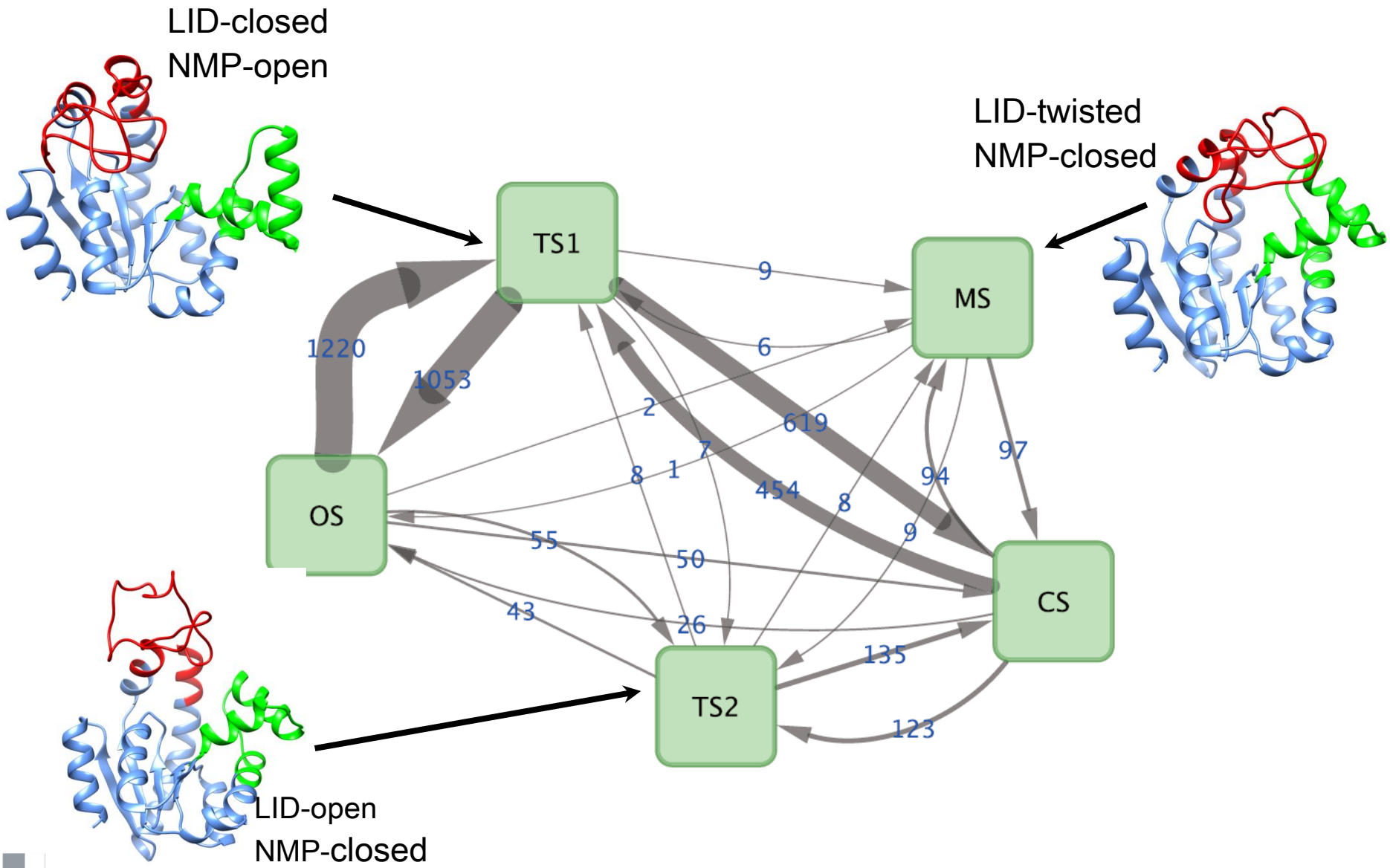


Table 2: **A comparison of top 16 highly fluctuating residues identified from the experiment and the simulation**

Type	Residues
NMR experiment	60, 103, 89, 129, 42, 144, 156, 138, *50*, 31, 55, 108, *189*, 124, 194, 2
$1 - S^2_{sim}$	**42**, 151, **131**, **147**, **143**, **156**, **127**, **57**, **102**, **139**, 7, **2**, **196**, **108**, **135**, 26
VAR	**42**, 7, **131**, **156**, 80, **102**, 151, 85, **144**, **30**, **57**, **108**, **2**, **195**, **124**, 26
DEV	**30**, 7, 80, **154**, **60**, **124**, **109**, **42**, 84, **130**, **138**, **196**, 12, **102**, **2**, **89**
Hinge regions†	29-30, 42-50, 59-61, 79-81, 110, 120, 158-161, 173-177

The residues are listed in the descending order of fluctuation. For the NMR data, more dynamics residues are represented by smaller S^2 values are less than 0.81 are shown. The residues adjacent to the peaks less than four residues apart are considered to be redundant, and not shown. For the simulation results, the residues, which agree with the experiment within the residue number difference of 3, are shown in bold.

$S^2_{sim} = \langle [3(\hat{\mu}_{\rm NH}(0)\hat{\mu}_{\rm NH}(t))^2 - 1]/2 \rangle$ with $t = 2$ ns ;

$\text{VAR} = \sqrt{\langle \phi^2 \rangle - \langle \phi \rangle^2} + \sqrt{\langle \psi^2 \rangle - \langle \psi \rangle^2}$;

$\text{DEV} = |\langle \phi \rangle_{TS1} - \langle \phi \rangle_{OS}| + |\langle \phi \rangle_{TS1} - \langle \phi \rangle_{CS}| + |\langle \psi \rangle_{TS1} - \langle \psi \rangle_{OS}| + |\langle \psi \rangle_{TS1} - \langle \psi \rangle_{CS}|$.

† The hinge regions were identified based on the variation of pseudo angles and pseudo dihedral angles between OS and CS.[29]

Acknowledgements

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

Direct band gap silicon: In-Ho Lee, *KRISS*

K. J. Chang, *KAIST*

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)