# Challenges and Solutions to Protein Structure Prediction in the Genomic Era

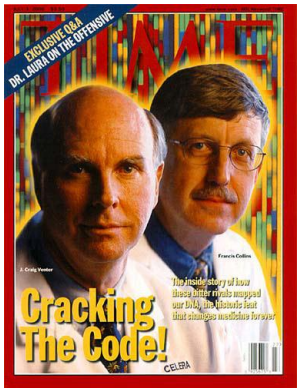

**Jianlin Jack Cheng, PhD**

**Computer Science Department**

**University of Missouri, Columbia**

**Oct. 5, 2011**

# The Genomic Era

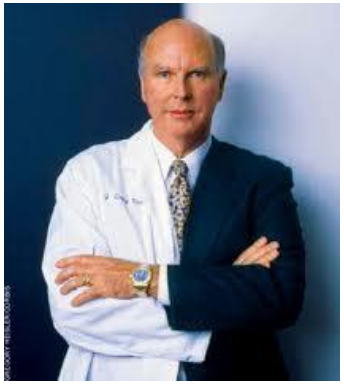




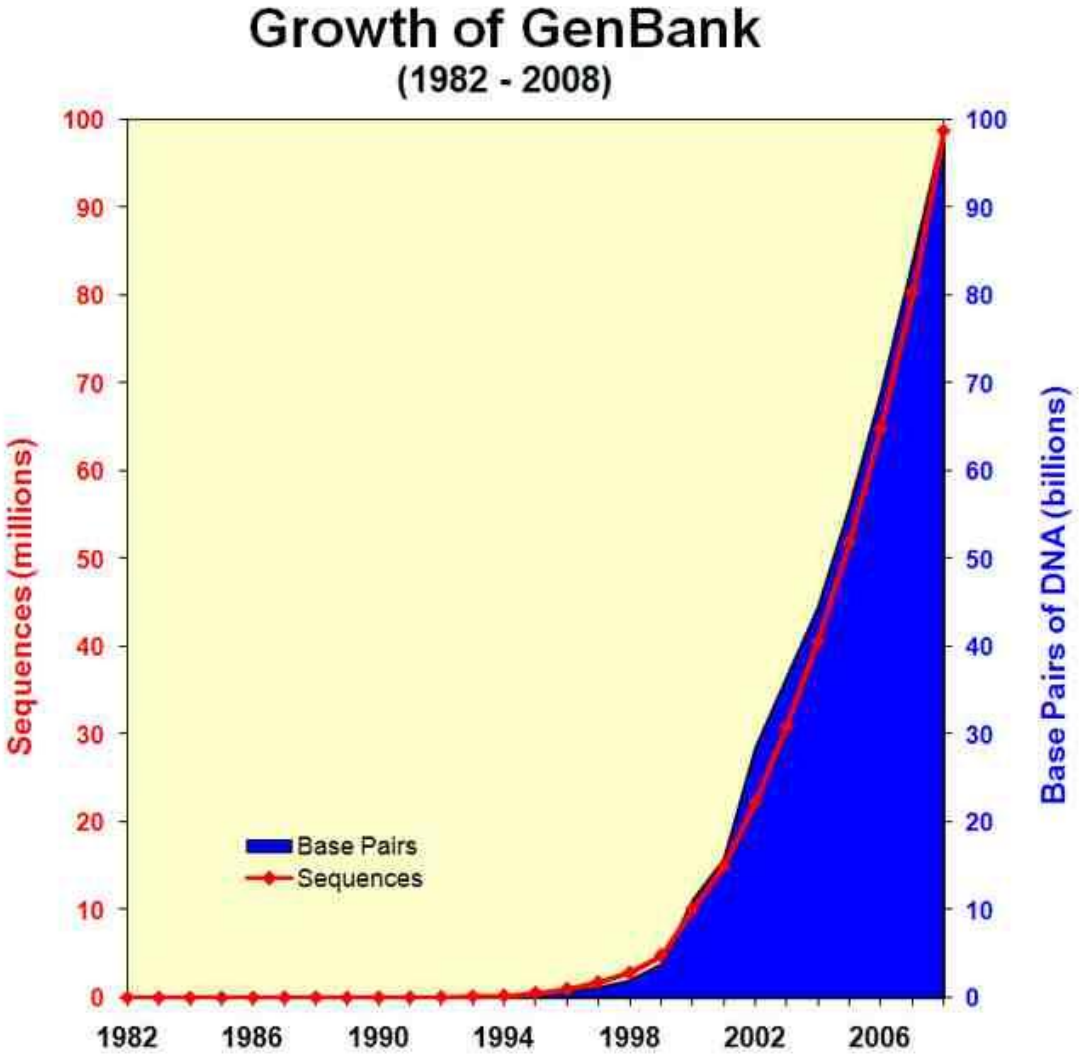

**Collins, Venter, Human Genome, 2000**


# NIH $1,000 Genome Project

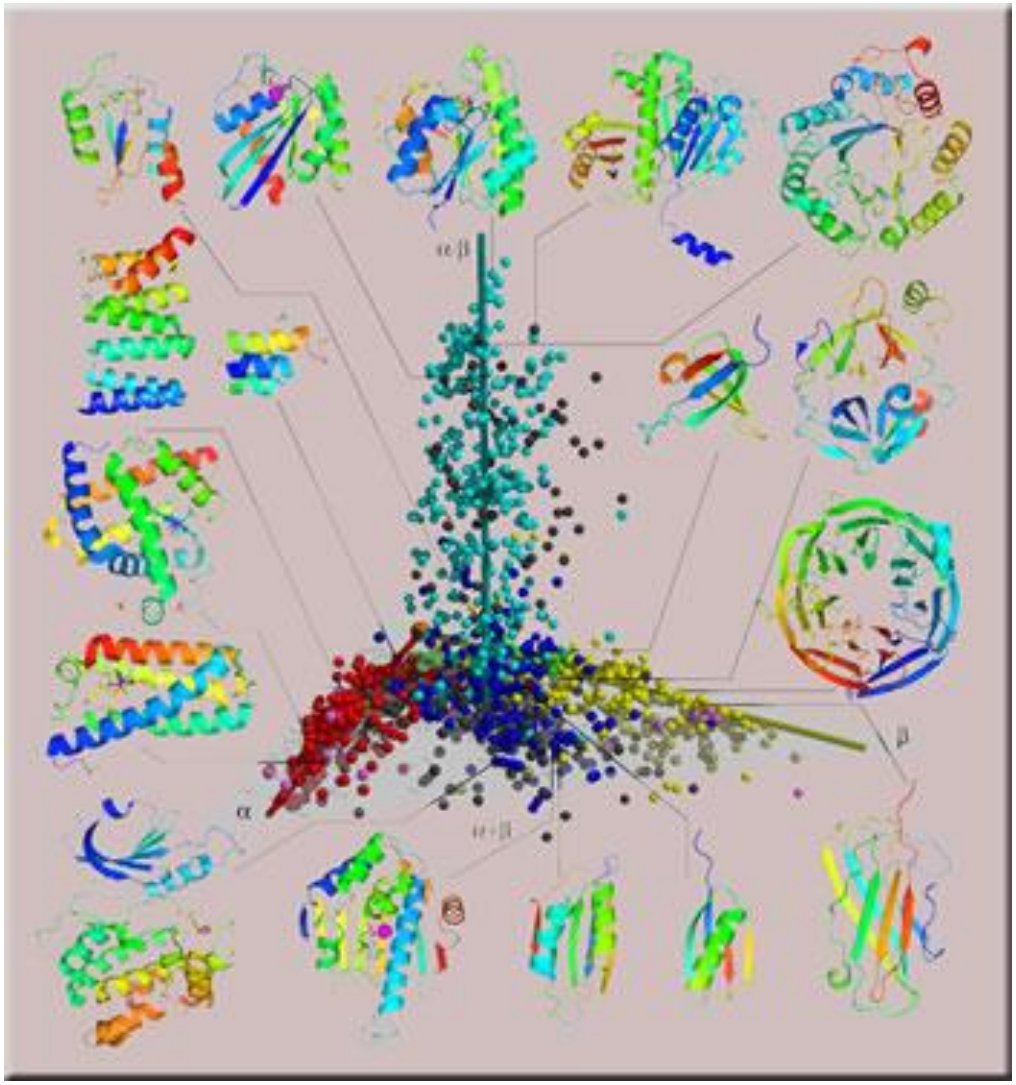
- Personal Computer in 1970s
- Personal Genome in 2010s



# Growth of Protein Sequences



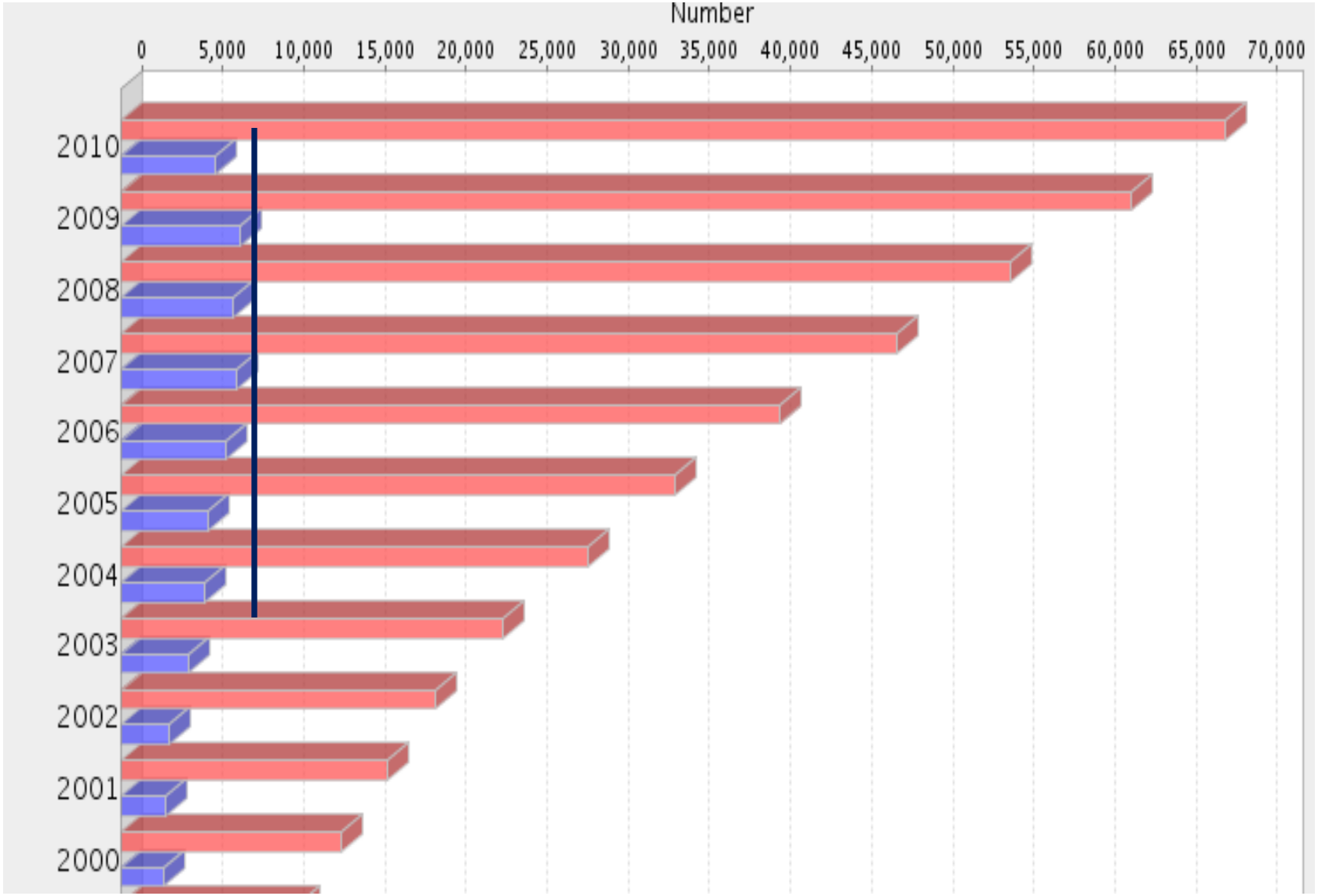
# NIH Structural Genomics Project



Chothia, Nature,1992

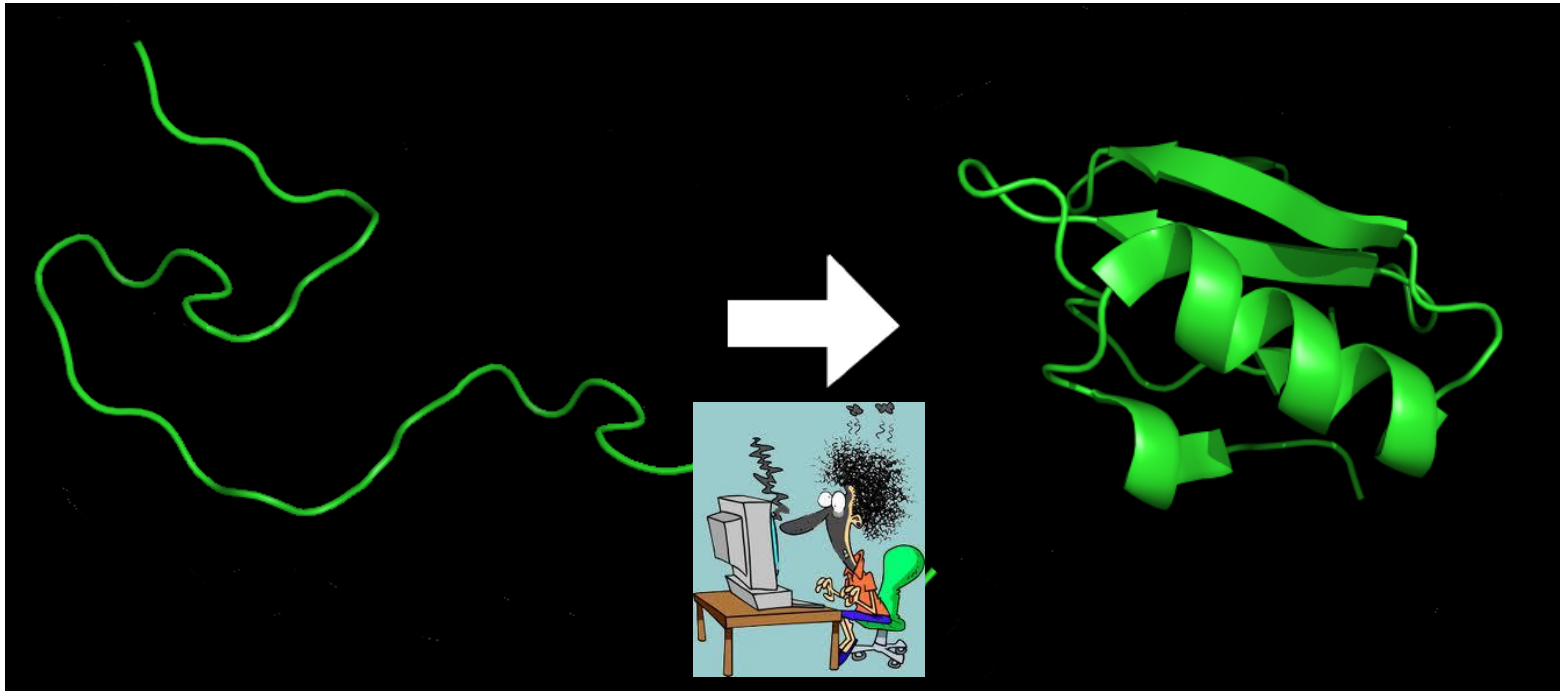
## Protein Sequence and Fold Space

# Growth of Protein Structures in PDB

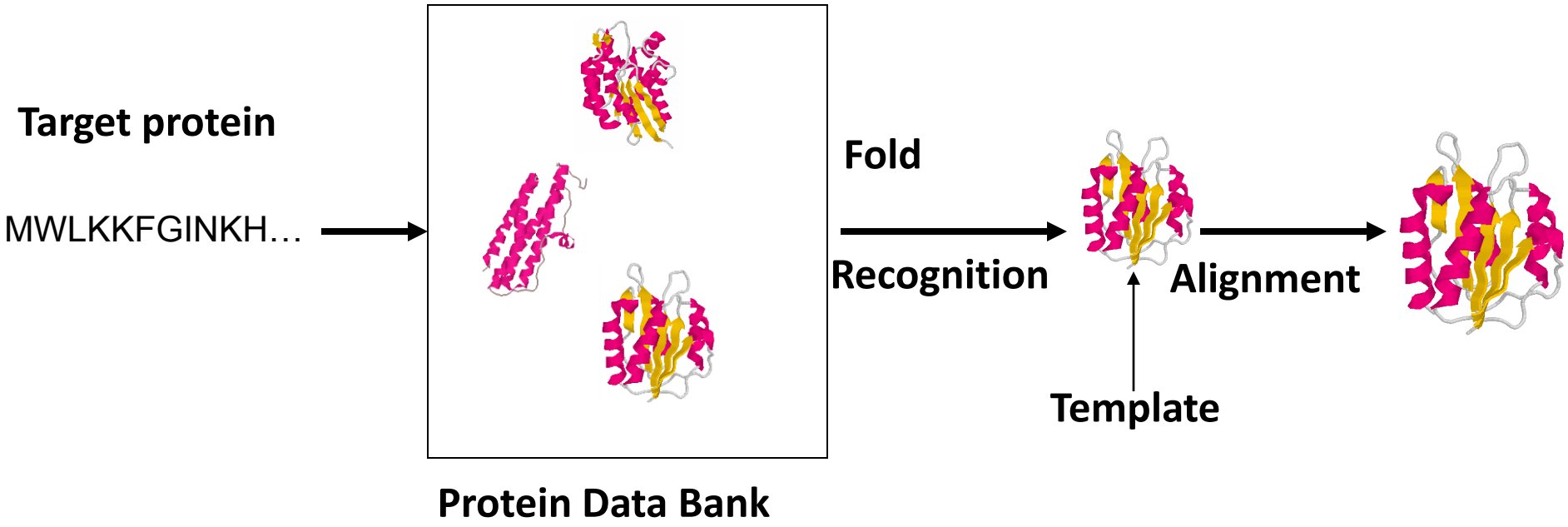


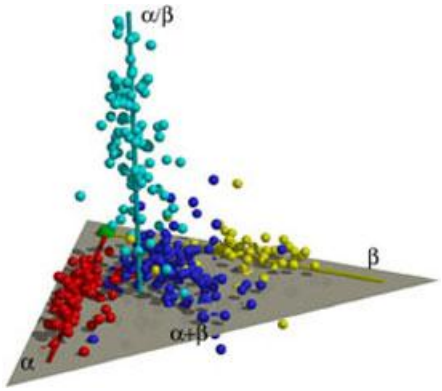
# Computational Protein Structure Folding / Prediction

**Structure = $f$ ( sequence) ? ←→ $E = MC^2$**



# Template-Based Approach




Protein Structure Space

TARGET

TEMPLATE

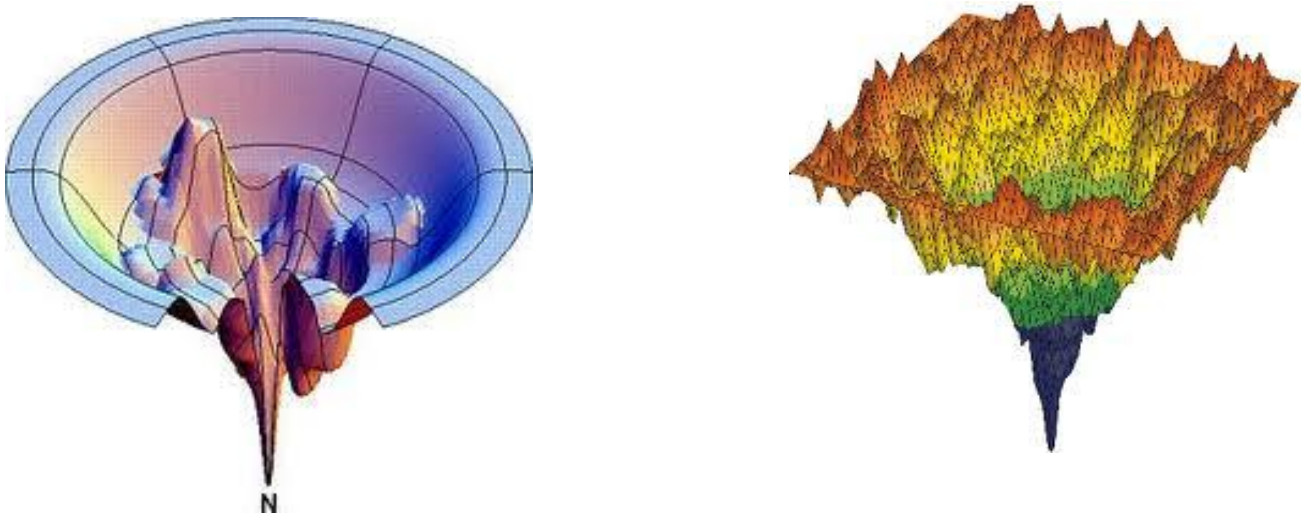
ASILPKRLFGNCEQTSDEGLK
IERTPLVPHISAQNVCLKIDD
VPERLIPERASFQWMNDK

ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIPE
MSVIPKRLYGNCEQTSEEAIRIEDSPIV---TADLVCLKIDEIPERLVGE

**Modeller**

A. Fisher, 2005

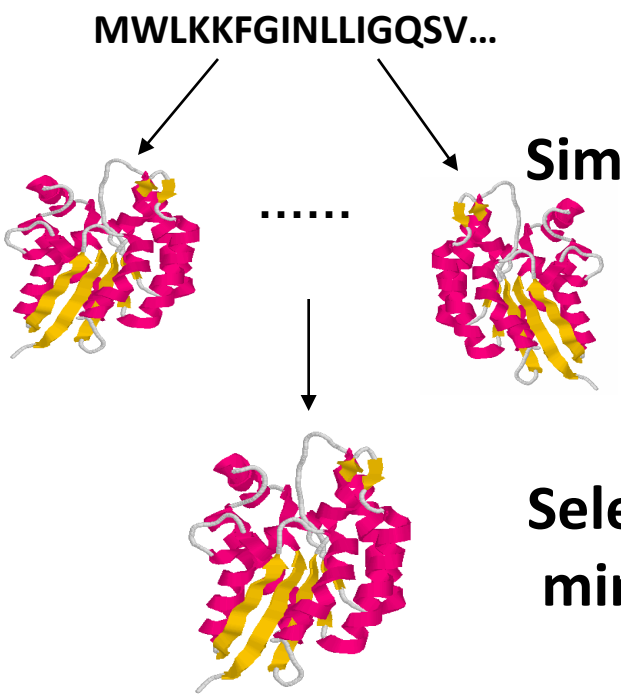
# Template-Free Protein Structure Prediction

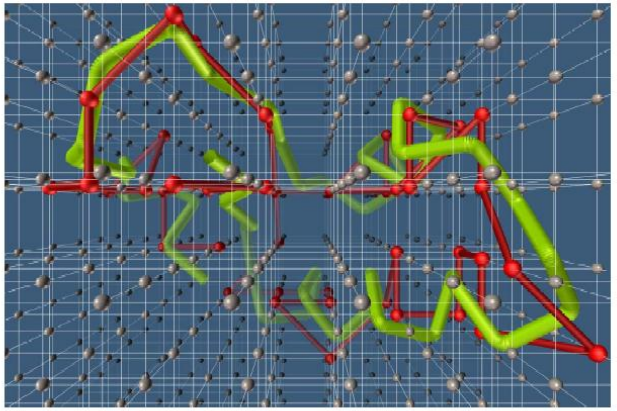



Dill & Chan, 1997

# Template-Free Approach

**Methods: molecular dynamics, fragment assembly, contact-based modeling**



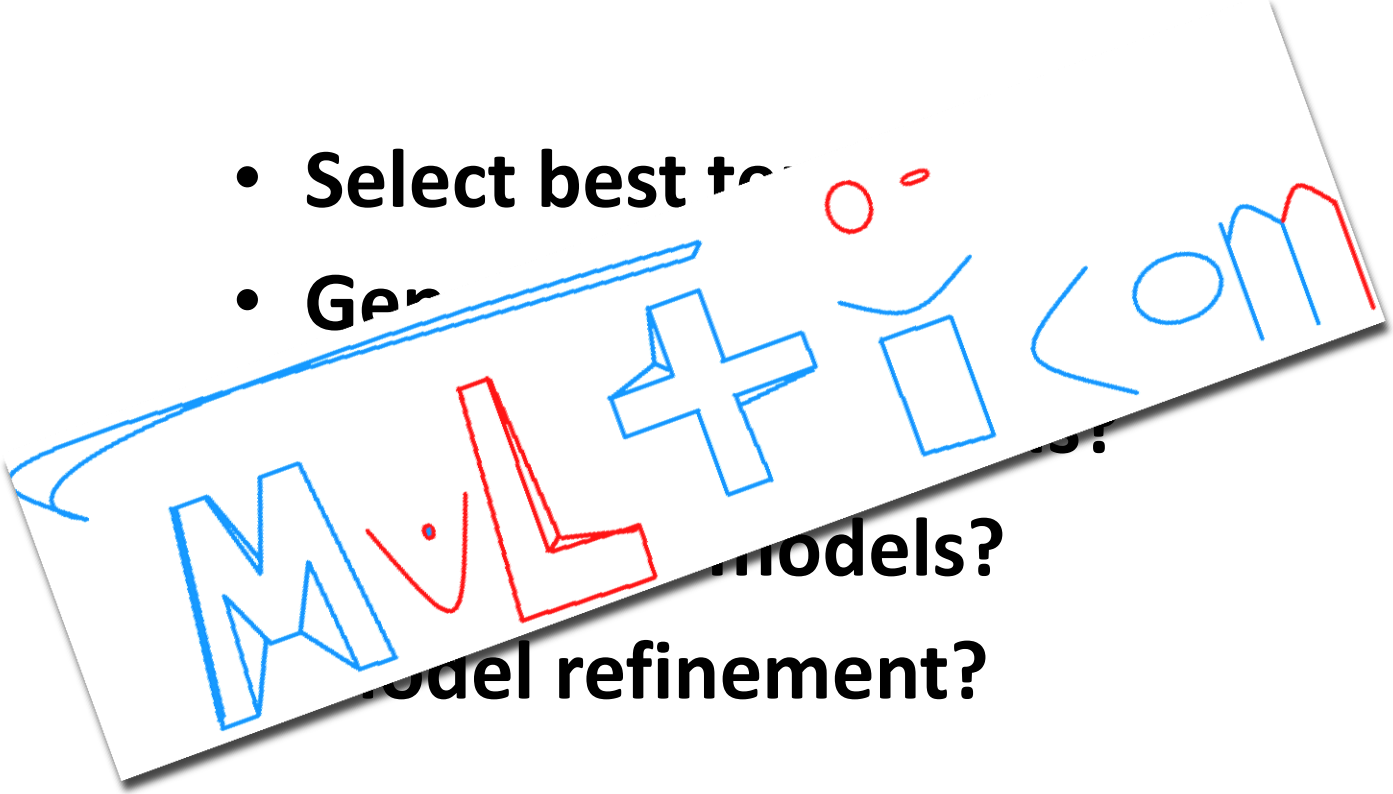



# Major Challenges in Protein Structure Prediction

- Select best templates?
- Generate best alignments?
- Generate best models?
- Select best models?
- Model refinement?

# Major Challenges in Protein Structure Prediction

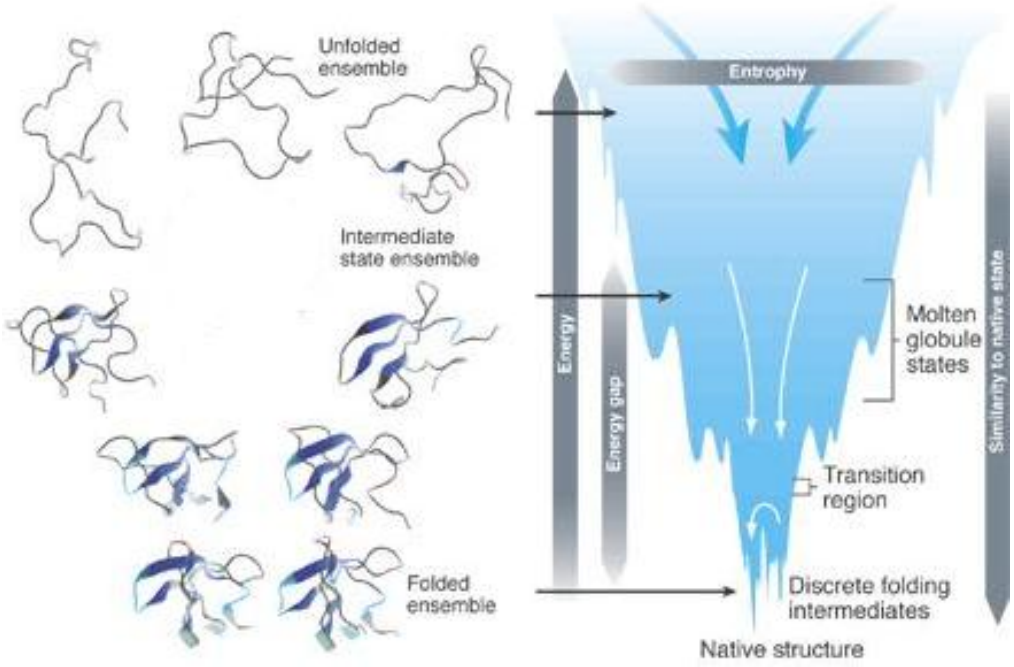
- Select best te
- Ge
- odels?
- odel refinement?



Z. Wang, J. Eickholt, J. Cheng, Bioinformatics, 2010

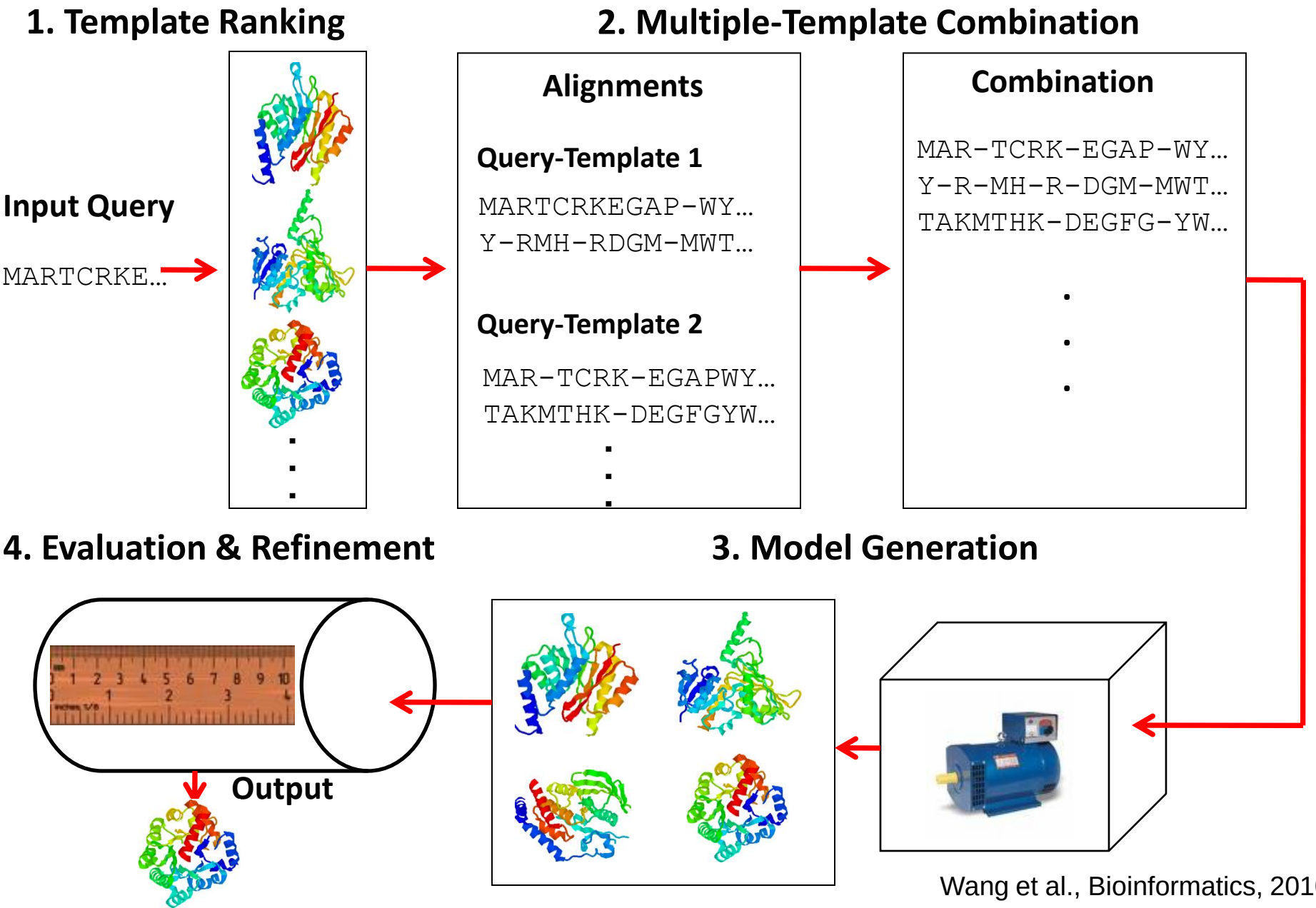
# Protein Conformation Ensemble

- ***P*(conformation) $\propto$ *P*(-energy)**
- **Conformation Distribution**
- **Maximum Likelihood & Maximum a Posterior**




Brooks et al., 2001

# A Unified Protein Structure Prediction Pipeline



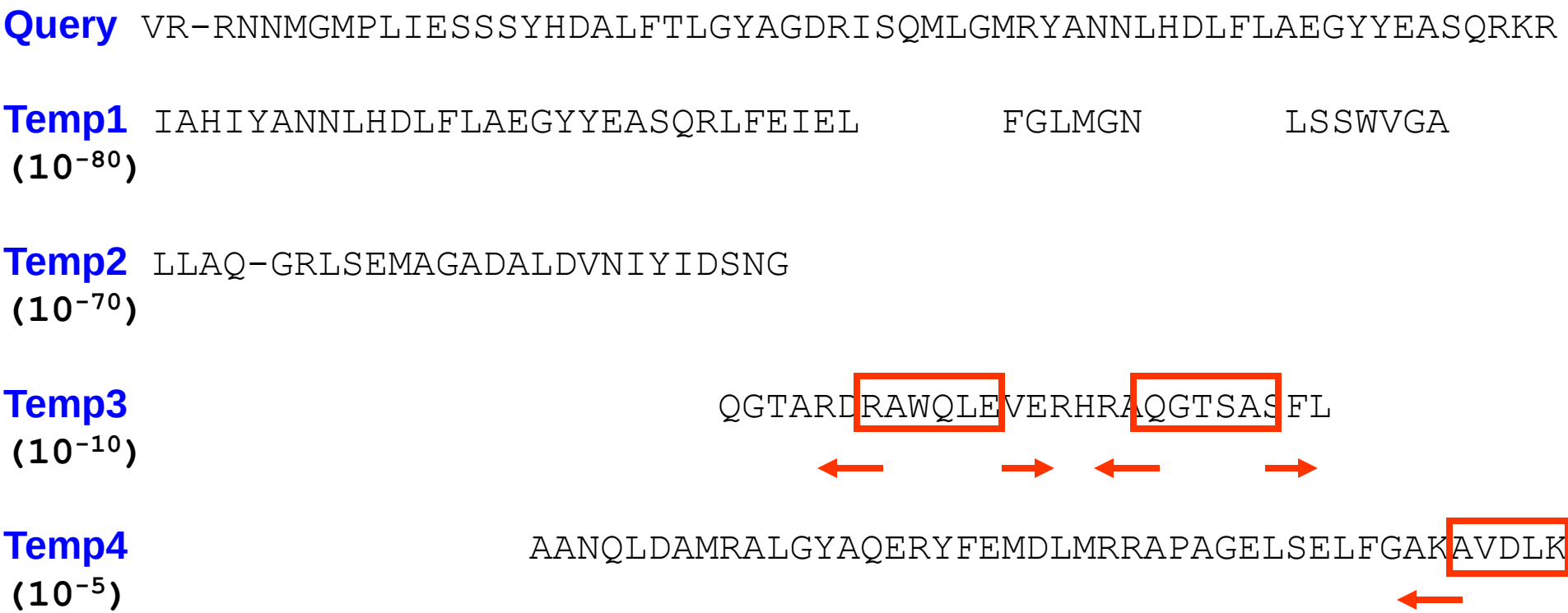

Wang et al., Bioinformatics, 2010

# Template Ranking / Fold Recognition

- **PSI-BLAST (sequence – profile)**
- **SAM (sequence – HMM)**
- **HMMer (sequence – HMM)**
- **Compass (profile – profile)**
- **HHSearch (HMM - HMM)**
- **PRC (HMM-HMM)**
- **FOLDpro (machine learning)**
- **MSACompro (profile-profile)**

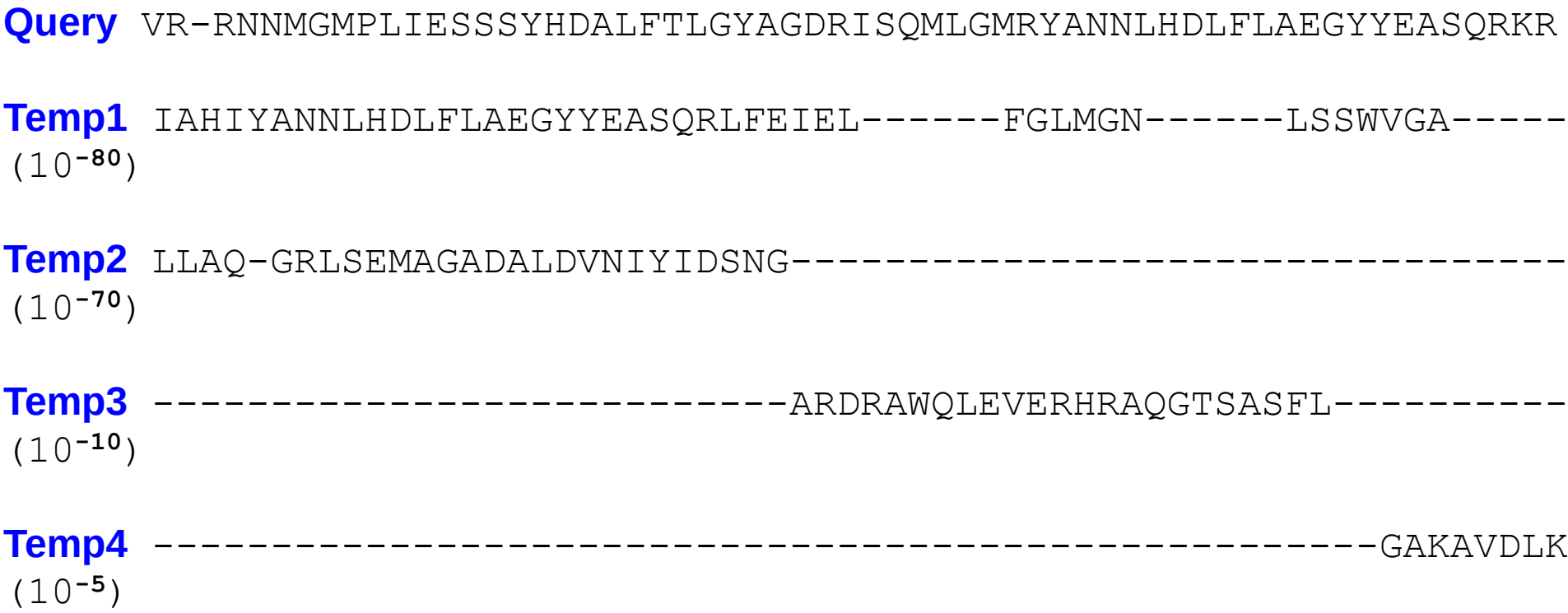
Cheng, Baldi, Bioinformatics, 2006
Deng, Cheng, Unpublished

# Multi-Template Combination in Template and Alignment Space

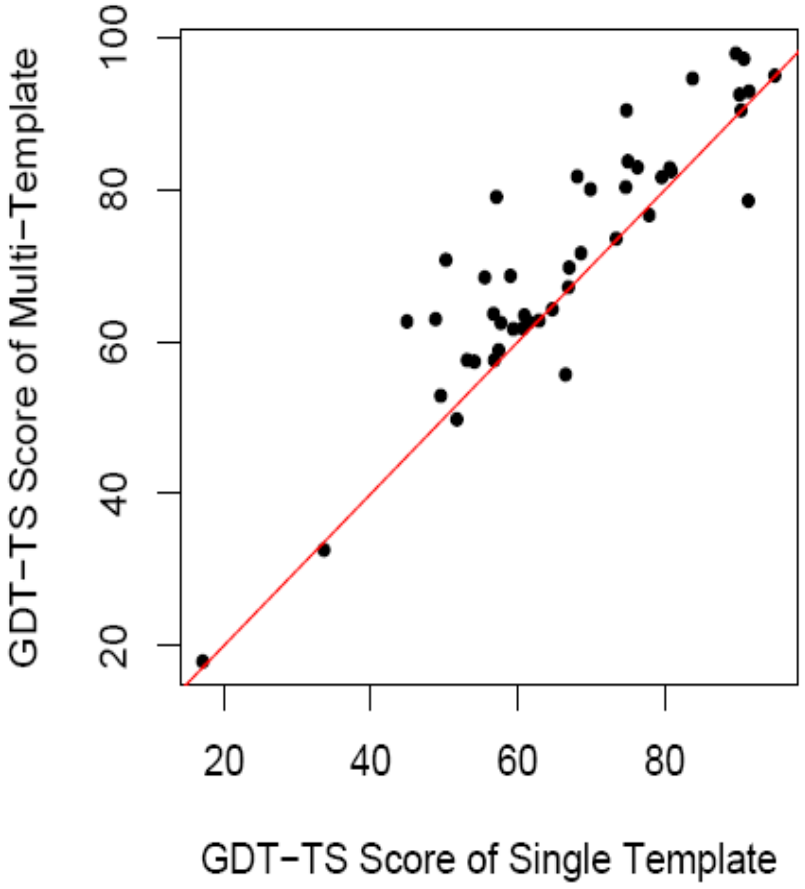



Cheng, BMC Structure Biology, 2008

# Multi-Template Combination in Template and Alignment Space



# Multi-Template VS Single-Top-Template

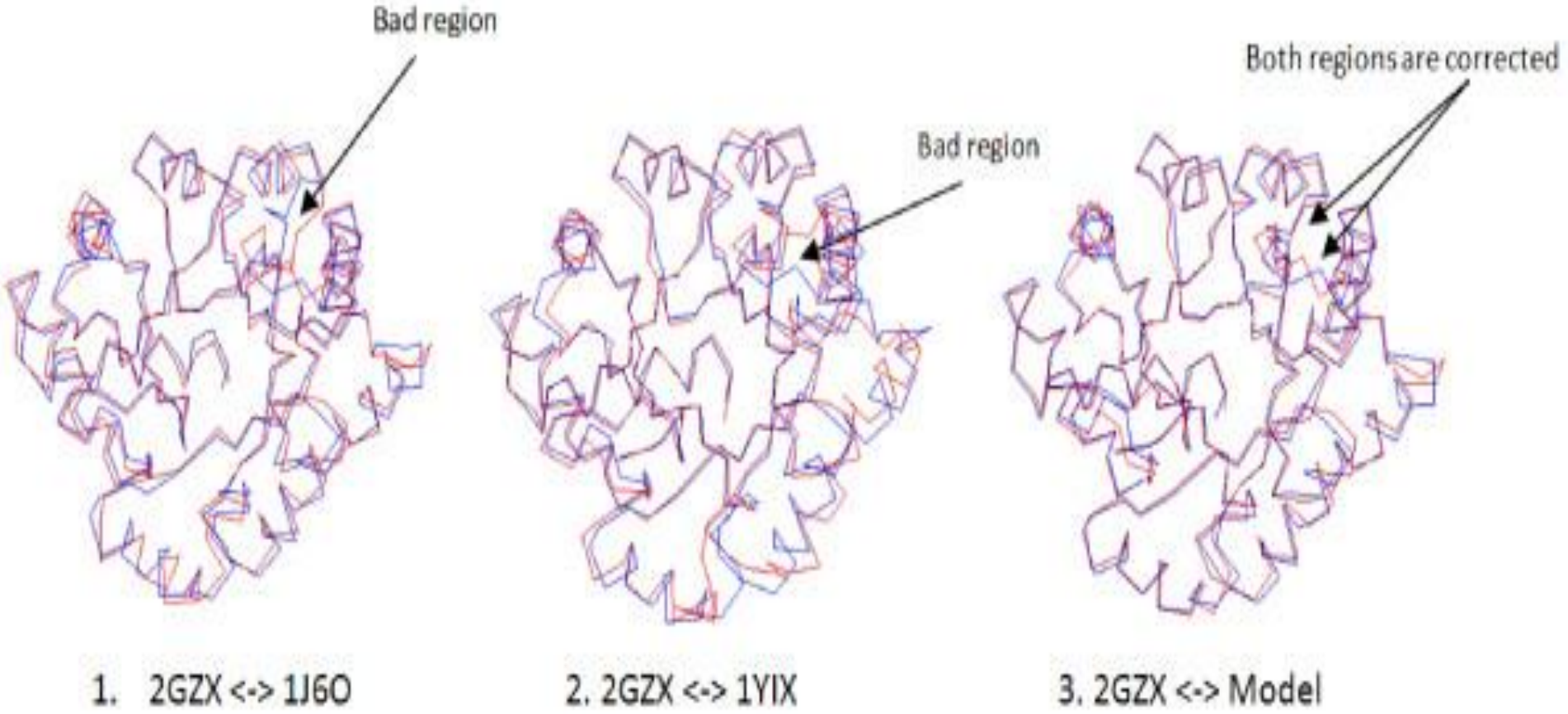



**Improve 38 / 45 targets**

**Improvement by 6.8%**

**P-value < $10^{-4}$**

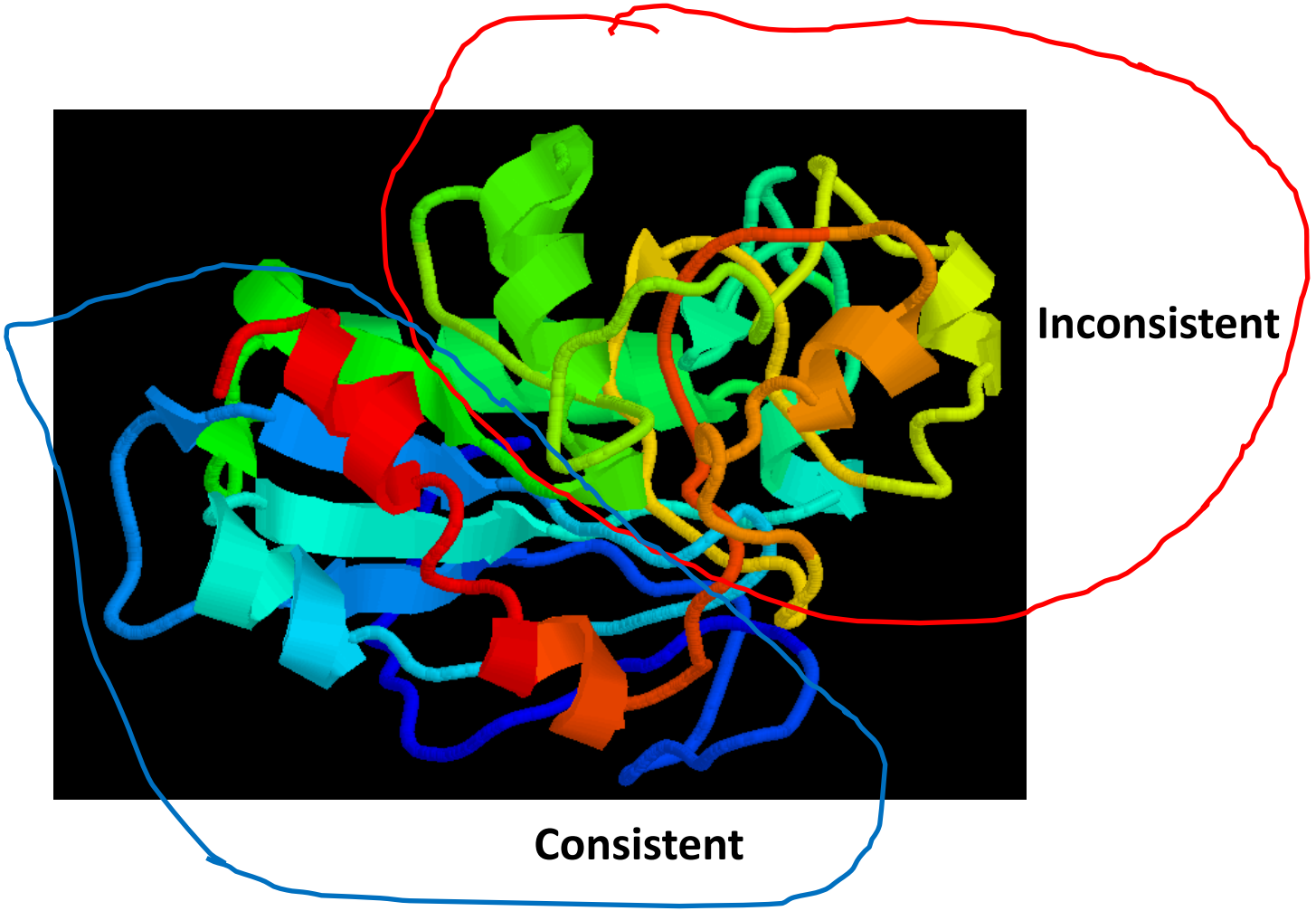
Cheng, BMC Structure Biology, 2008




Cheng, BMC Structure Biology, 2008

**Advantage: reduce variance of modeling**

# Problem – Atom Clashes




**Combine 49 templates – CASP8 target T0413**

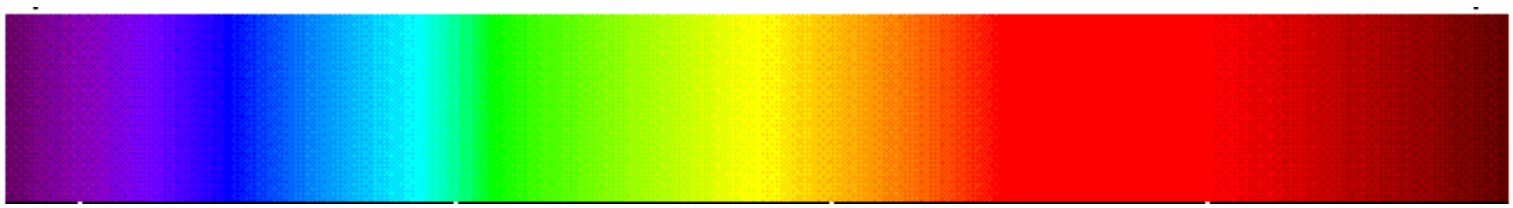
# Structure Comparison-Guided Multi-Template Combination (CASP9)

- **Check the structural consistency between query-template alignments**
- **Only combine query-template alignments that are structurally consistent**
- **Completely removed atom clashes in CASP9**

Cheng et al., 2010

# Combination of Template-Free and Template-Based Modeling for Model Generation

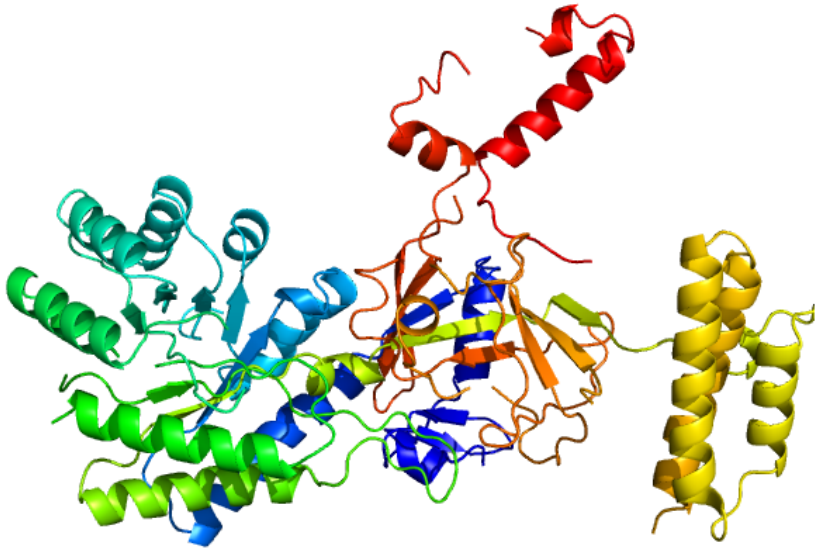
100% TBM 50% TBM+50%FM 100% FM



Protein Modeling Spectrum

# Region Decomposition

A)



B)

**Template: 1TWIA**

```
-------MLGNDTVEIK-DGRFFI---DGYDA-IELAEK---------FGTPLYVMSEEQIKINYNRYI
EAFKRWEEETG--KEFIVAYAYKANANLAITRLLAKLGC----GADVVSGGELYIAKLSNVPSK----K
IVFNGNC-KTKEEIIMGIE---ANIRA-FNVDSISELILINETAKE-LGETANVAFRINPNVNPKTHPK
ISTGLKKNKFGLDVESGIAMKAIKMALE--MEYV-NVVGVHCHIGSQLTDISPFIEETRKVMDFVVELK
E-----E-GI-EIEDVNLGGGLGIPYYKDKQI----PTQKDLADAIINIMLKYKD--KVEMPNLILEPG
RSLVATAGYLLGKVHHIKETPVT----------------------------------------------
----------------------------------------------KWVMIDAGMNDMMRP-AMYE
AY-HHIINCK----VKN----EKEVVSIAGGLCESSDVFGRDR-------ELD-KVEVGD---VLAIFD
VGAYGISMAN-NYNARGRPRMVLTS--KKG-V--FLIRERETYADLIAK----------------------
----------------------------
```

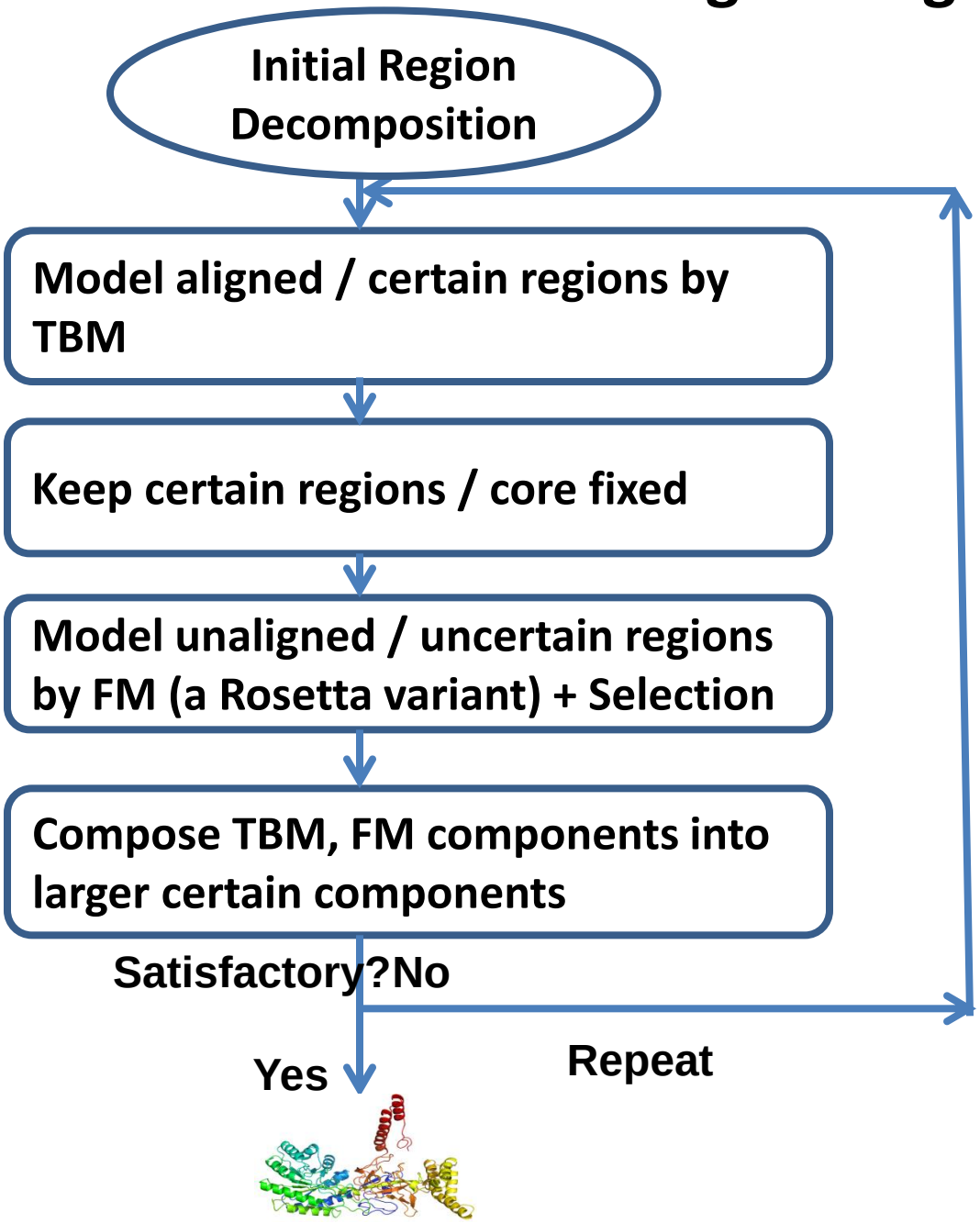
**Query: T0547**

Domain 3

Domain 4

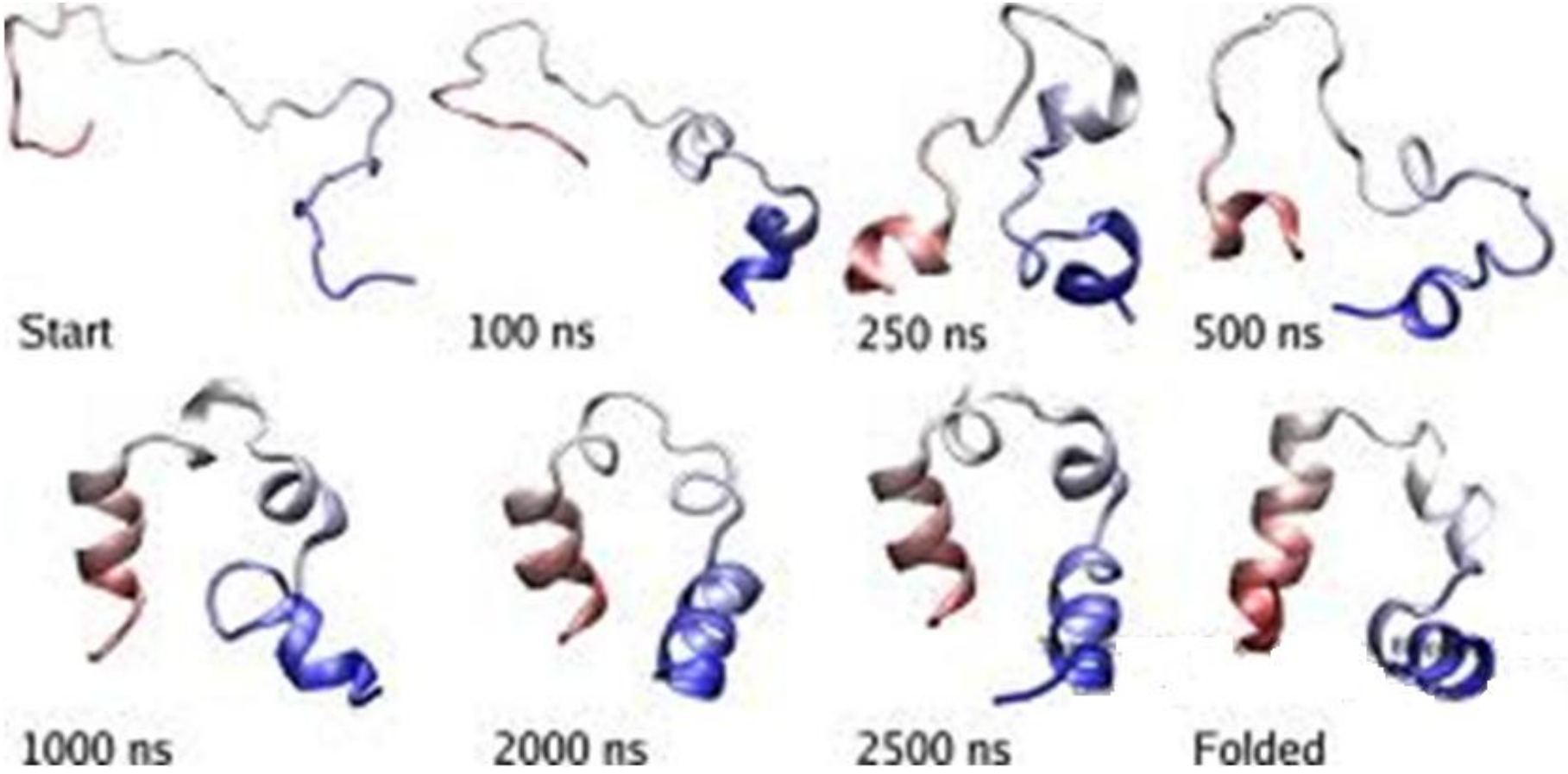
```
MMDYGIDIWGNENFIIK-NGKVCINYEKKPAI-IDIVKELR----DDGYKGPLLLRFPHLIQKQIENIY
GNFNKARKEFGYKGGFNAVYPLKVNQYPGFVKNLVKLGKDYNYGLEAGSKAELLLAMAYNNEGA----P
ITVNG-F-KDRELINIGFIAAEMGHNITLTIEGLNEVEAIIDIAKERFKPKPNIGLRVRLHSAGVGI-W
AKSGGINSKFGLTSTE--LIEAVNLLKE--NKLLEQFTMIHFHLGSQITEIHPLKKALNEAGNIYTELR
K-----M-GAKNLKAINLGGGLAVEYSQEKNEKSRNYTLREYANDVVETLKNIAEQKKDLEPDIFIESG
RFVAANHAVLIAPVLELFSQEYAENKLILKKQNPKLID-ELYDLYKSI--KPSNALEYLHDSIDHLESI
LTLFDLGYVDLQDRSNAEILTHLITKKAILLLGDKQNPADLLAIQDEVQERYLVNFSLFQSMPDFWGLE
QN-FPIMPLD----RLD--EEPTRSASIWDITCDSDGEISYSKD---KPLFLH-DVDVEKENYFLGFFL
VGAYQEVLGM-KHNLFTHPTEAIISINEKG-YEVEGIIEAQSILDTLEDLDYDIHAIMDILNERISNSK
LVNDKQKKHILGELYLFLNDNGYLKSIGV
```

# Recursive Protein Modeling – Integrate TBM and FM



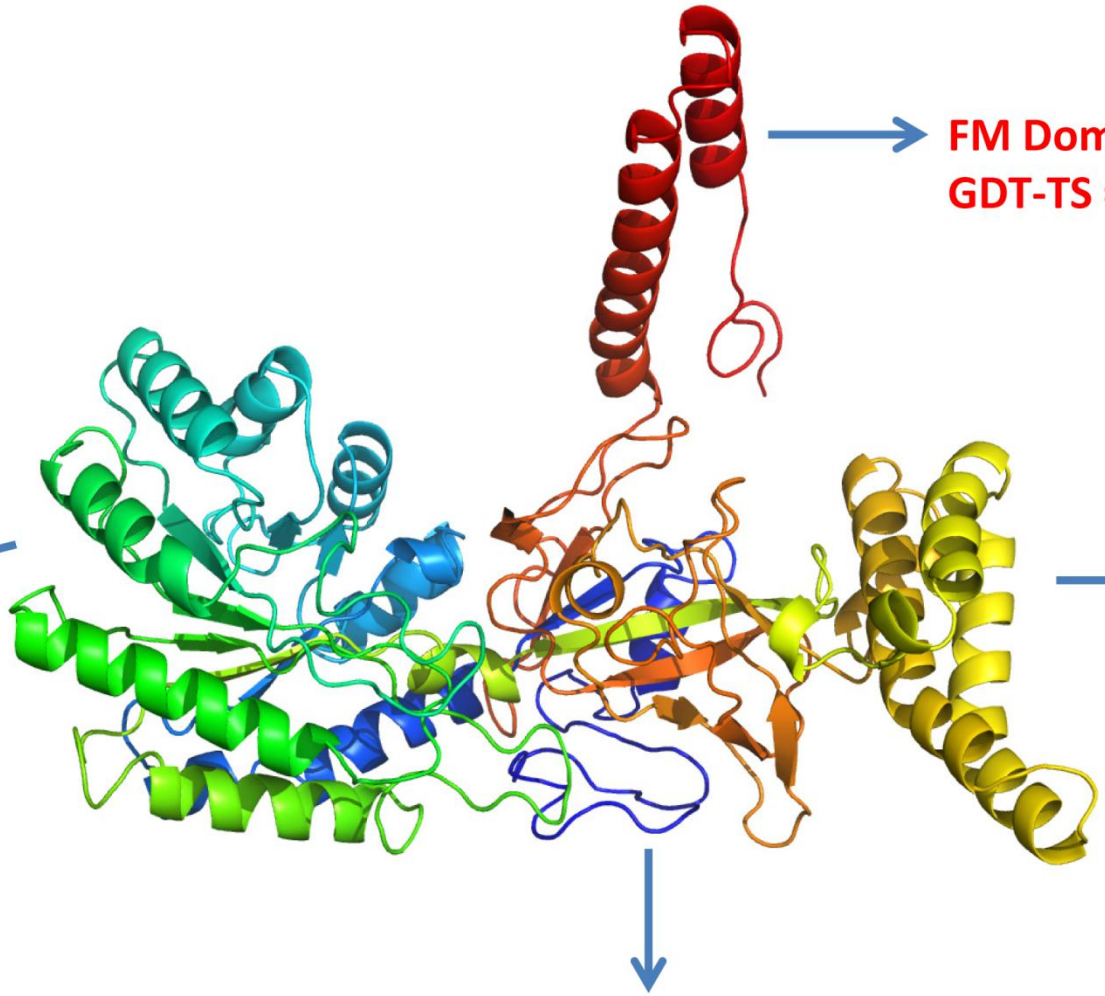

Divide & Conquer

# Recursive Modeling Mimics Protein Folding Cascade




ks.uiuc.edu

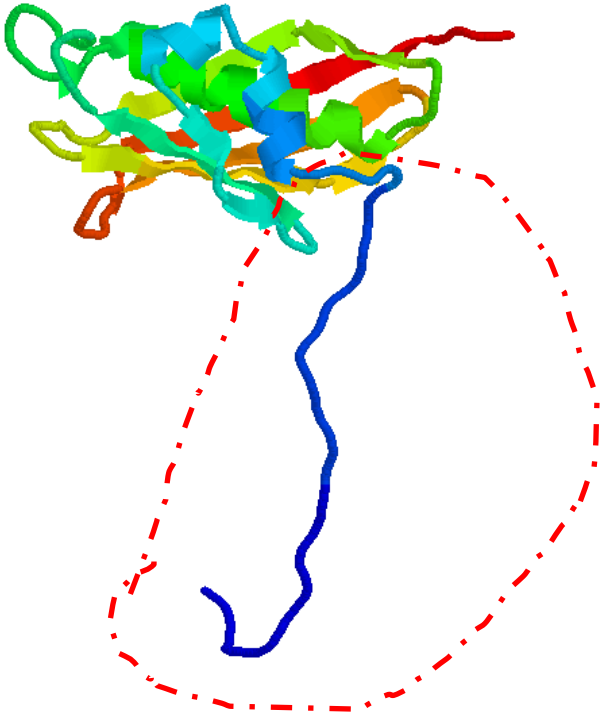
# Domain Level Integration (CASP9)



# Core-Constrained Tail Refinement

```
>P1;1VI8B
structureX:1VI8B: 1: : 146 : : : :
--------------------------SLIWKRKITLEALNAMGEGNMVGFLDIRFEHIGDDTLEATMPVDSRTKQPFGLLHGGASVVL
AESIGSVAGYLCTEGEQKVVGLEINAHVRSAREGRVRGVCKPLHLGSRHQVWQIEIFDEKGRLCCSSRLTTAILEGGSHHHHH*


>P1;Query
  : : : : : : : : :
MDKRLQQDRIVDKMERFLSTANEEEKDVLSSIVDGLLAKQERRYATYLASLTQIESQEREDGRFEVRLPIGPLVNNPLNMVHGGITATL
LDTAMGQMVNRQLPDGQSAVTSELNIHYVKPGMGTYLRAVASIVHQGKQRIVVEGKVYTDQGETVAMGTGSFFVLRSRG-----*
```

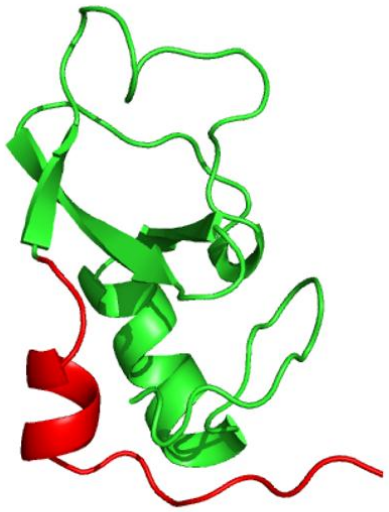


# Template-Based + Template-Free & Recursive Modeling (CASP9)

A)

Before tail refinement
GDT-TS = 0.64

B)



After tail refinement
GDT-TS = 0.73

C)

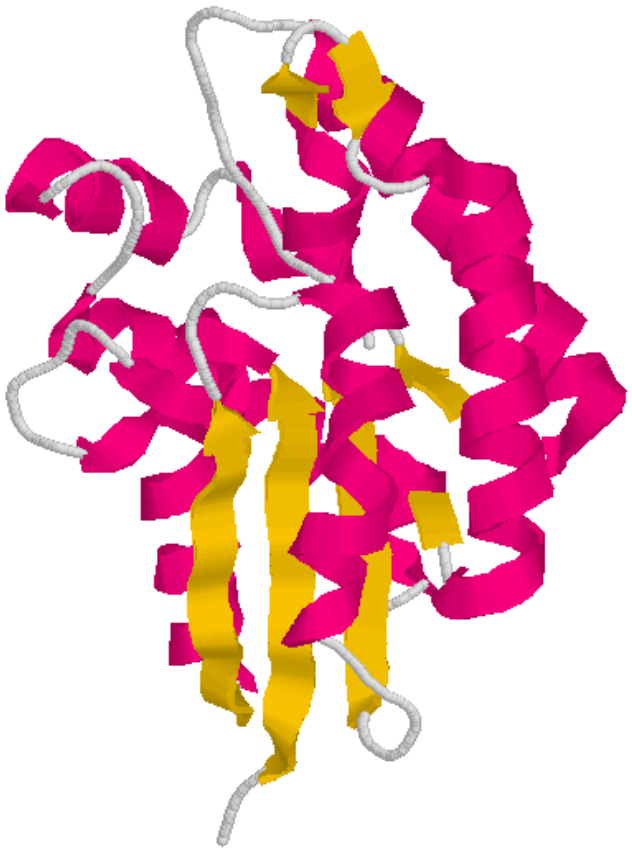
Superposition
Green: model, Blue: structure

# Improve Template-Free Modeling Using Contact Restraints

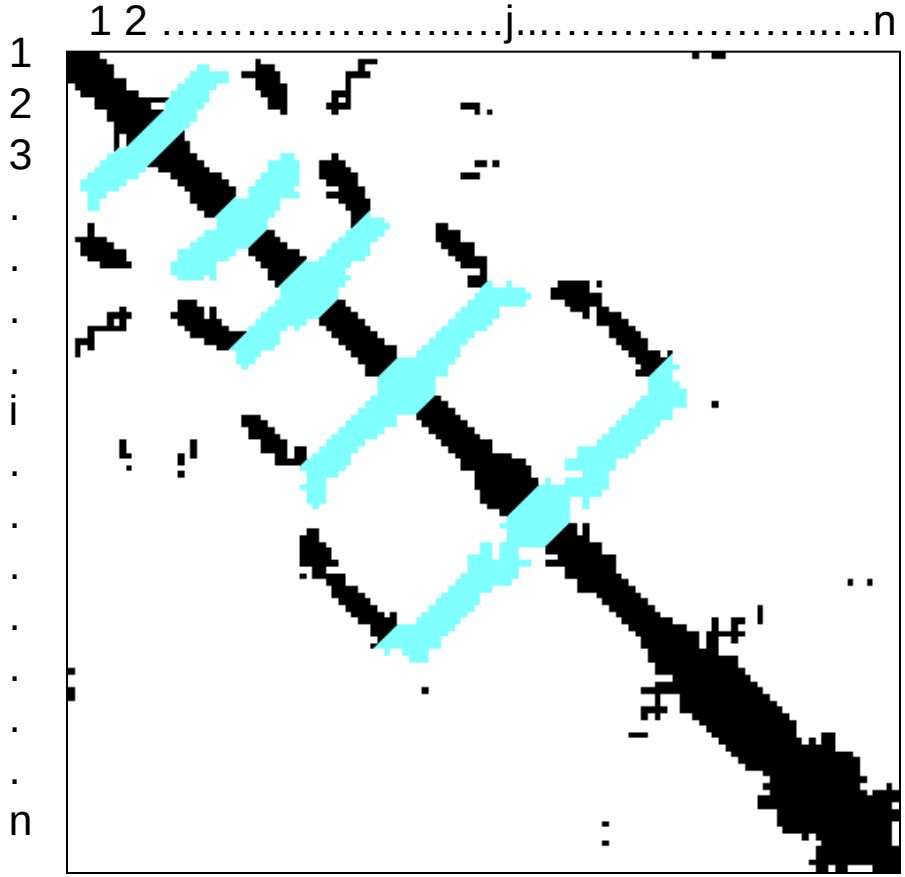
- Fragment assembly (e.g. Rosetta)
- Conformations with good secondary structures and compactness, but often wrong topology
- Huge bottleneck (30% proteins)

# Contact Map Prediction

## 3D Structure



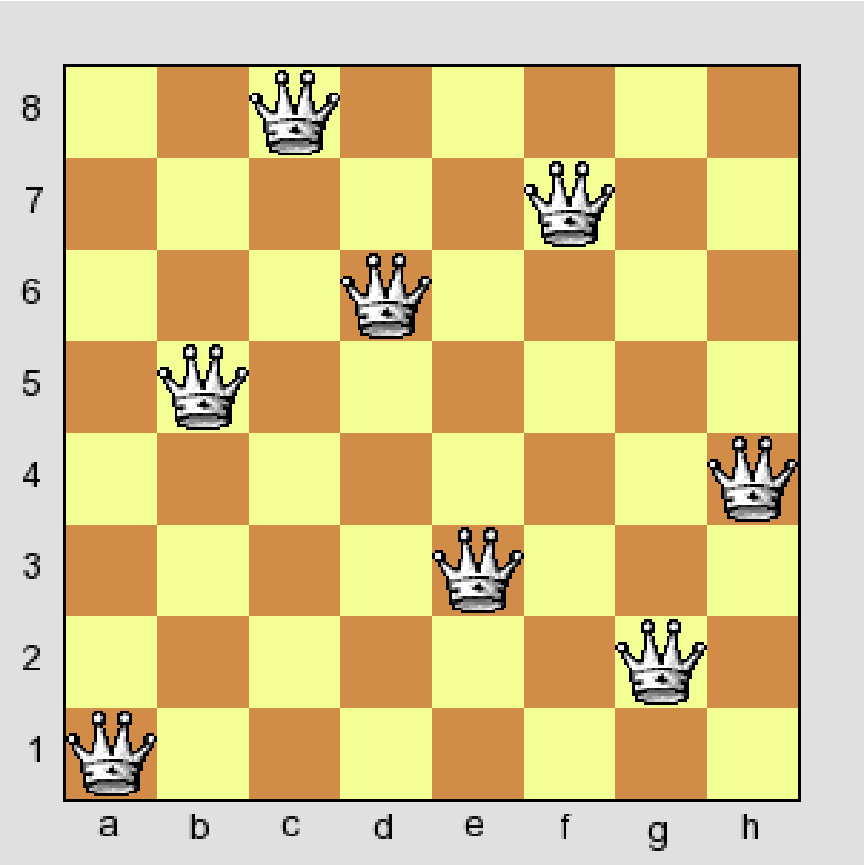
## 2D Contact Map




Cheng, Randall, Sweredoski, Baldi, 2005
Cheng and Baldi, 2007.
Tegge et al., 2009

# Contact-Based Protein Structure Reconstruction

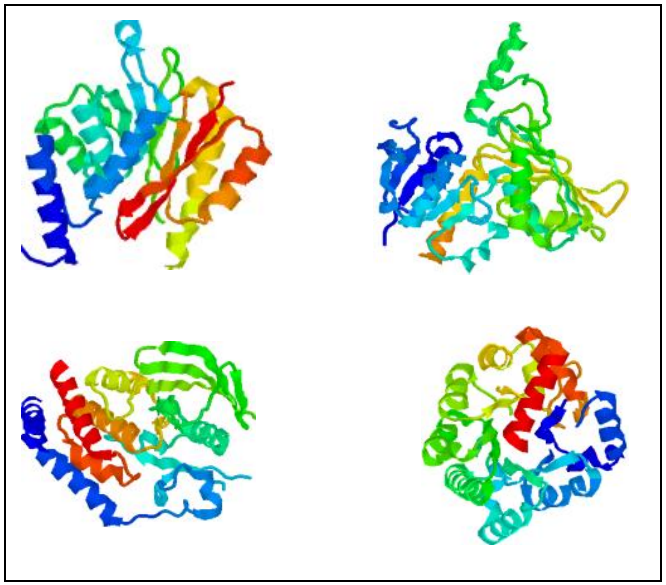
- A constrained optimization problem
- A small number of contacts L / 10
- 30% accuracy




Eight Queens Problem

# Contact Prediction

- **2D Neural Networks** (CASP7, 8, 9)
- **Support Vector Machine** (CASP7, 8, 9)
- **A Conformation Ensemble Approach** (CASP9)



**Contact Voting**

Eickholt, Wang, Cheng, BMC Bioinformatics, 2011

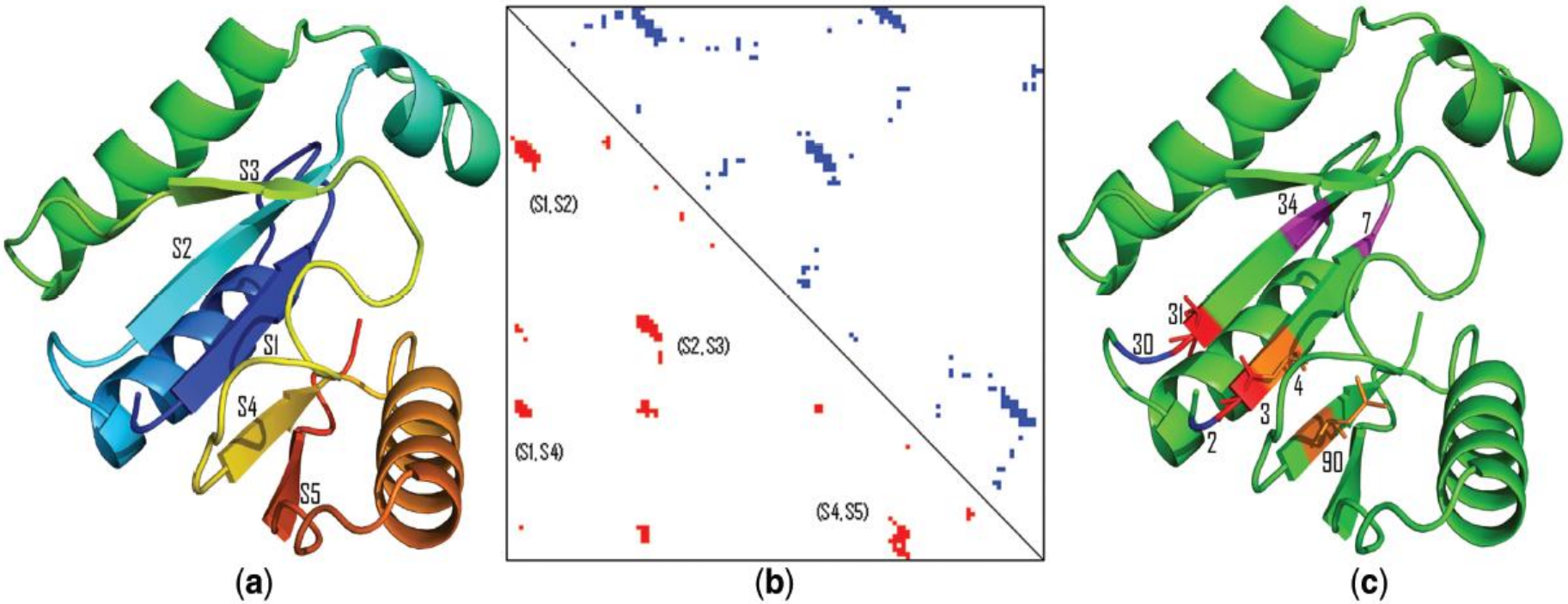
# CASP9 Contact Predictions


| Contact Predictors | >= 12 | >=24 |
|---|---|---|
| **Ensemble** | **.34** | **.30** |
| **Ensemble (±1)** | **.55** | **.48** |
| Zhang-Server | .28 | .23 |
| RosettaVote | .27 | .20 |

**Top L/5 contacts, on 26 CASP9 *ab initio* domains**

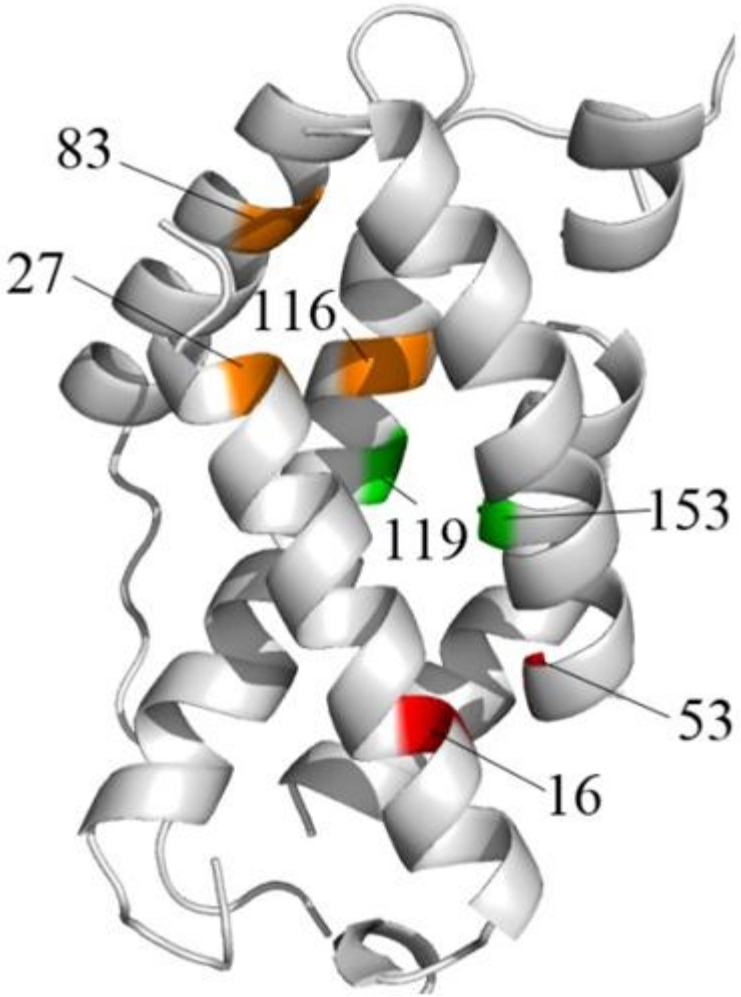
Eickholt, Wang, Cheng, BMC Bioinformatics, 2011

# A CASP8 Example: T0507




Tegge et al., Nucleic Acids Research, 2009

# A CASP9 Example




Eickholt, Wang, Cheng, BMC Bioinformatics, 2011

# Model Evaluation

- **Single model approach**
- **Pair-wise model assessment**
- **Hybrid approach**

Wang et al., Proteins, 2009
Cheng et al., Proteins, 2009
Wang et al., Bioinformatics, 2011

# ModelEvaluator

3D Model

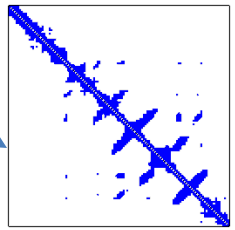
Comparison

Secondary Structure

EEEECCEEEHHHHHHHHHHHHEEEECCEEEHHHH

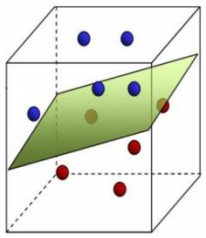
Relative Solvent Accessibility

eeee-----eeeee---------eeeee------eeeee---eeeeeeee



Contact Map

Achieve correlation ~0.7



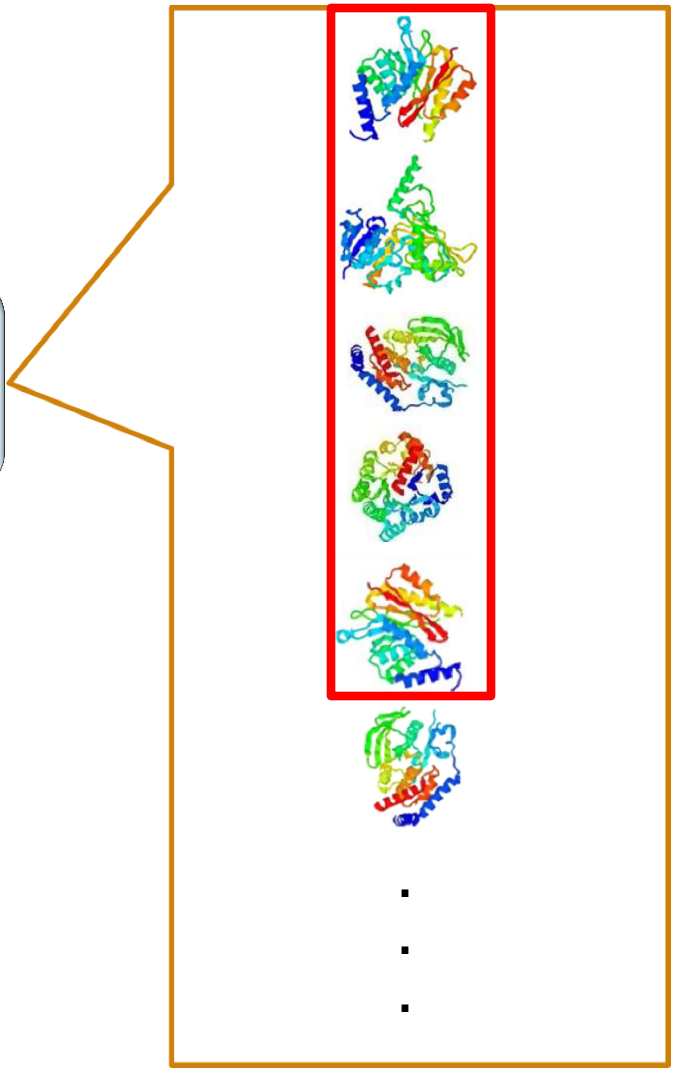
SVM

Predicted quality score

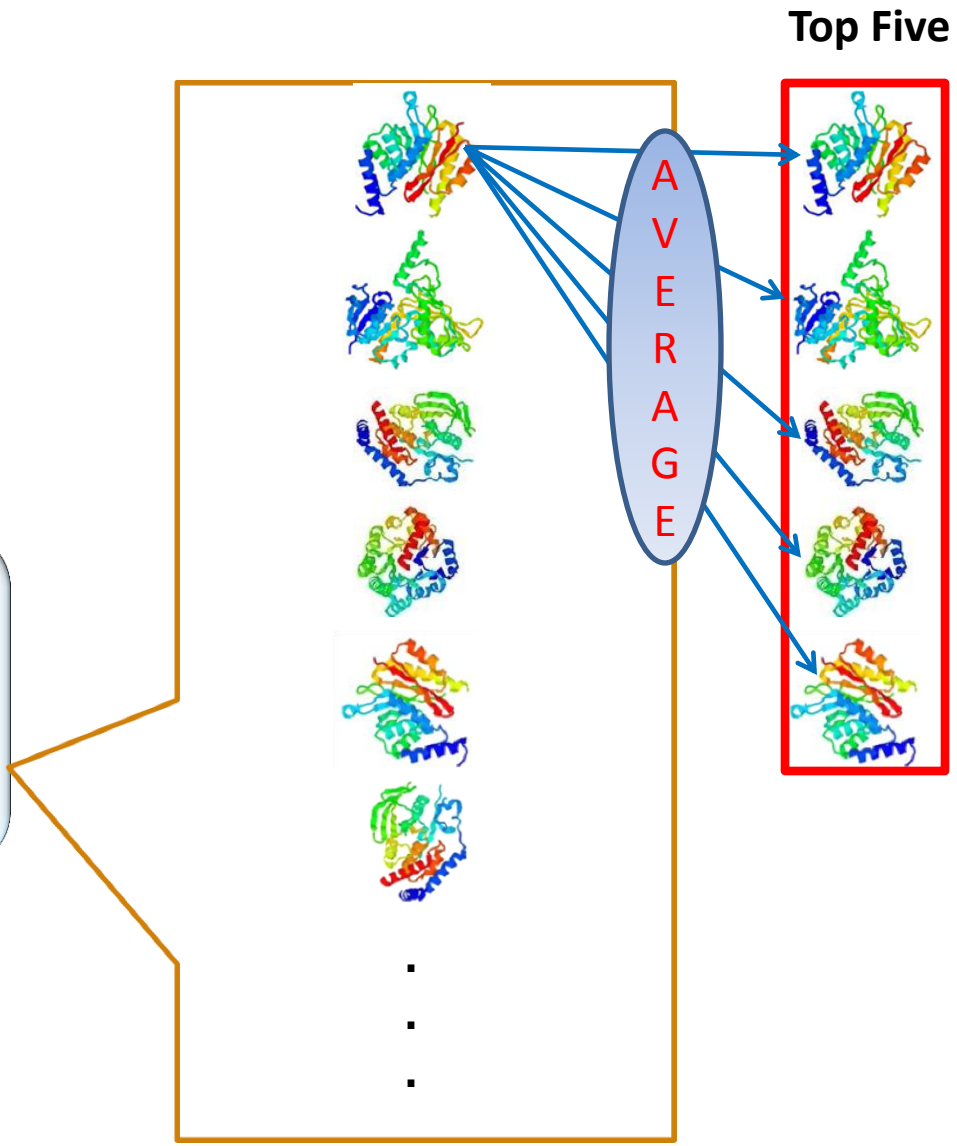
Wang et al., Proteins, 2009

# Model Quality Evaluation

Select top 5 ranked models as references



# Model Quality Assessment



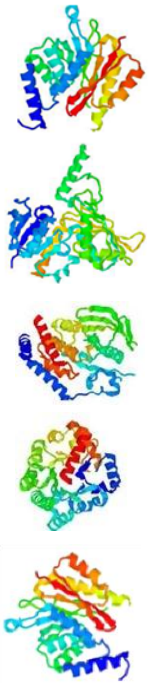

Compare each model with reference models
Average global quality

Re-rank models (+10%)

Cheng et al., Proteins, 2009

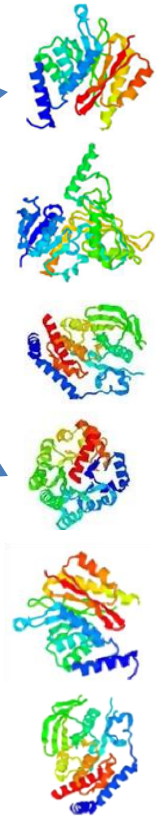
# Global-Local Model Combination

Model ranking



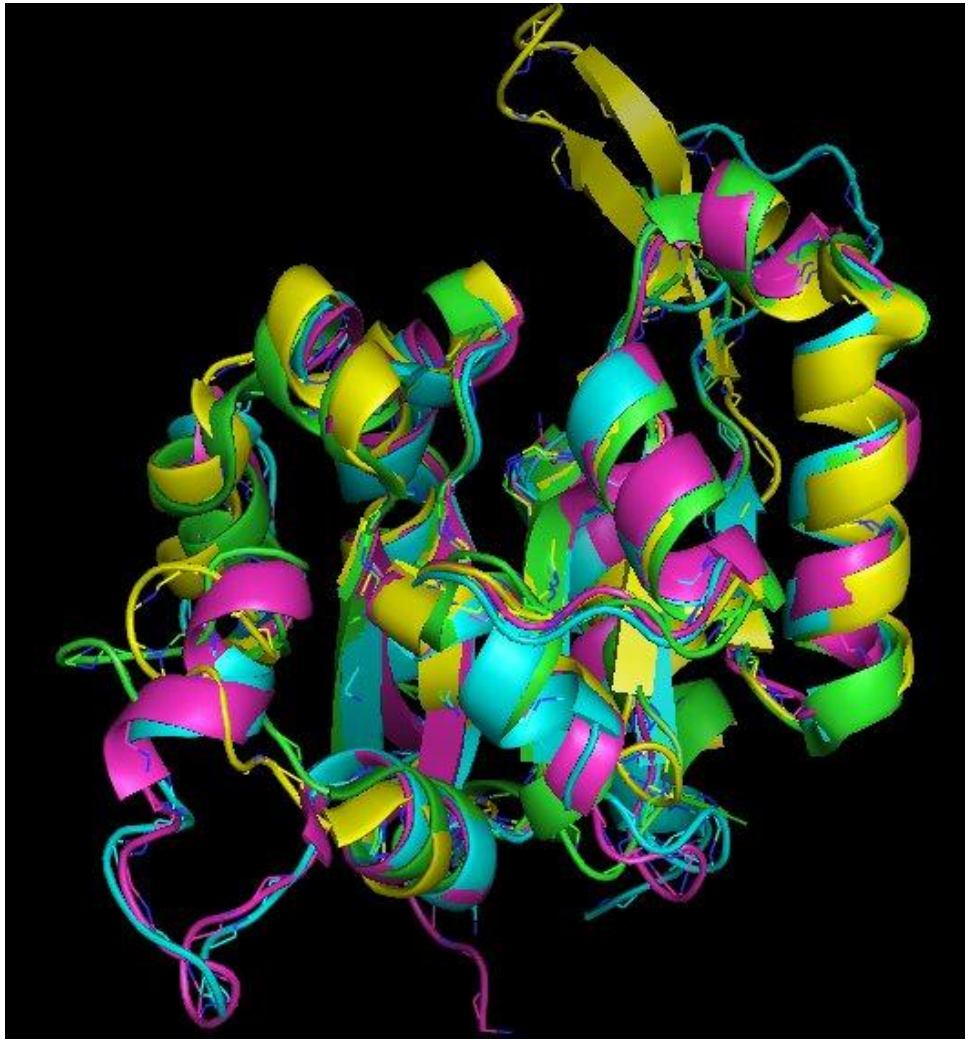
Select top 5 models
as seed models

Structure comparison

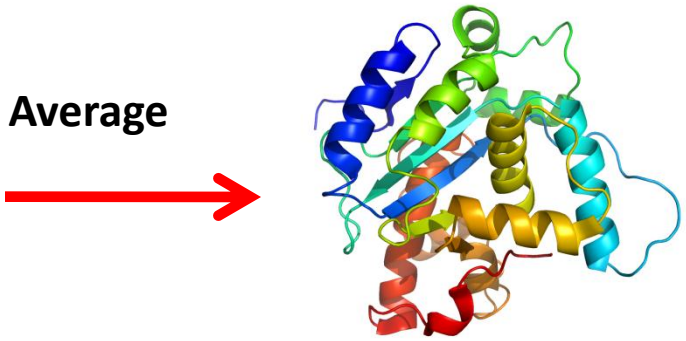


Identify similar models
or fragments

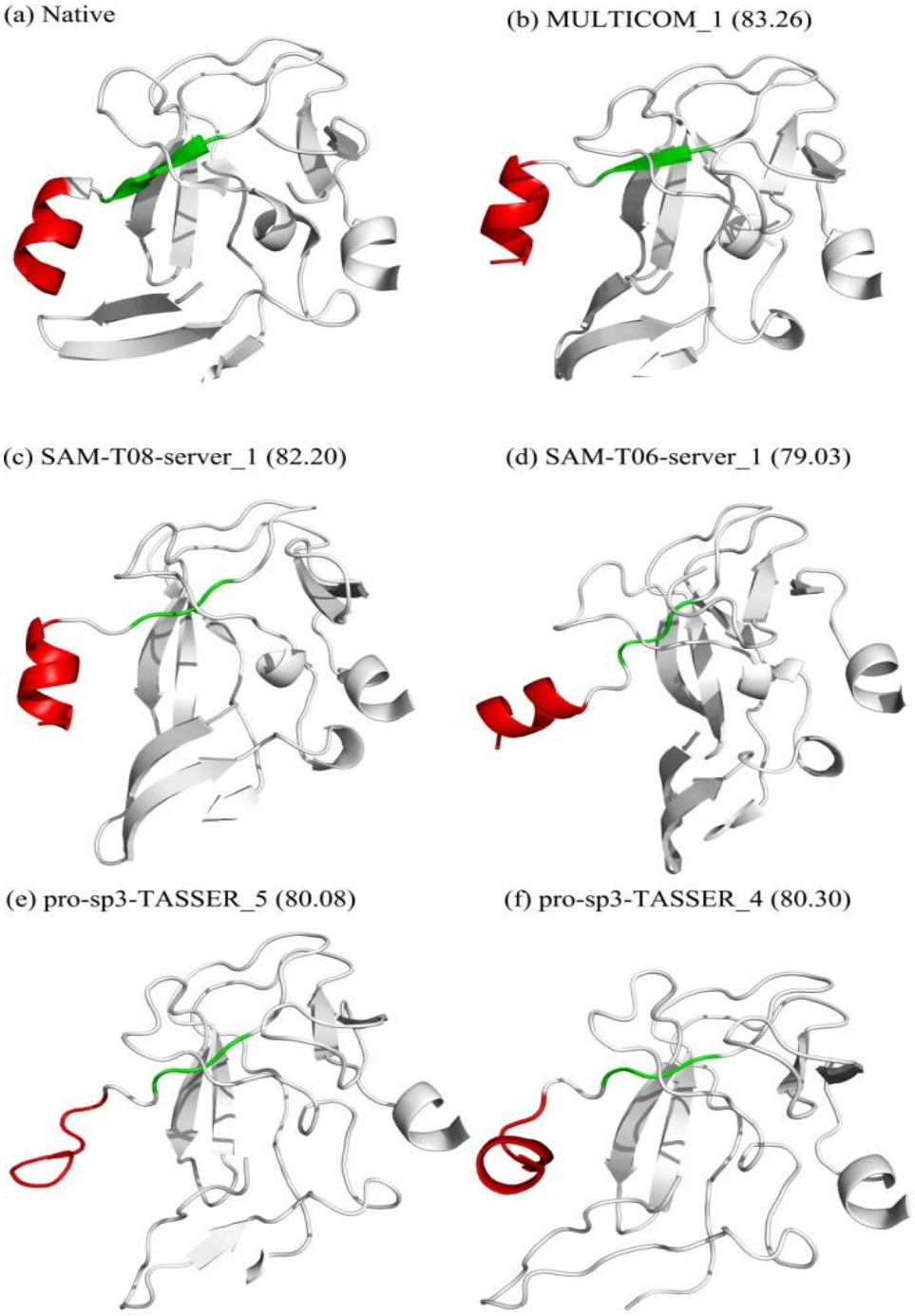
# Model Combination and Averaging



Average



**Advantage: reduce variance of modeling while increasing fitness**





Wang et al., Bioinformatics, 2010

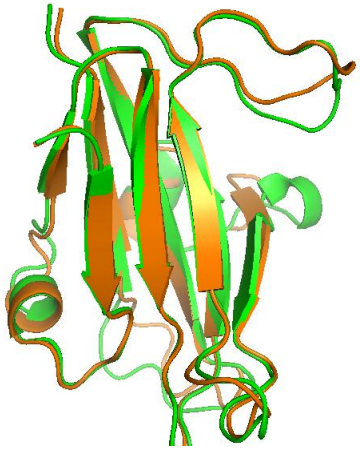
# CASP9 Top 20 Servers

| # | GR # | GR name | Domains Count | SUM Z-score (GDT_TS) | AVG Z-score (GDT_TS) | AVG GDT_TS |
|---|---|---|---|---|---|---|
| 1. | 380 | QUARK | 147 | 115.788 | 0.788 | 62.675 |
| 2. | 428 | Zhang-Server | 147 | 113.242 | 0.770 | 62.765 |
| 3. | 077 | RaptorX-MSA | 147 | 103.270 | 0.703 | 61.774 |
| 4. | 286 | RaptorX | 147 | 103.010 | 0.701 | 61.731 |
| 5. | 276 | RaptorX-Boost | 147 | 99.845 | 0.679 | 61.453 |
| 6. | 449 | HHpredB | 147 | 93.104 | 0.633 | 59.528 |
| 7. | 453 | HHpredA | 147 | 93.104 | 0.633 | 59.528 |
| 8. | 346 | HHpredC | 147 | 91.821 | 0.625 | 59.361 |
| 9. | 452 | Seok-server | 147 | 89.542 | 0.609 | 60.158 |
| 10. | 002 | MULTICOM-CLUSTER | 147 | 88.944 | 0.605 | 59.987 |
| 11. | 321 | BAKER-ROSETTASERVER | 145 | 87.240 | 0.602 | 58.768 |
| 12. | 119 | MULTICOM-REFINE | 147 | 86.441 | 0.588 | 59.519 |
| 13. | 215 | MULTICOM-NOVEL | 147 | 82.825 | 0.563 | 59.371 |
| 14. | 236 | gws | 145 | 82.645 | 0.570 | 58.931 |
| 15. | 457 | chunk-TASSER | 147 | 82.609 | 0.562 | 58.846 |
| 16. | 174 | Phyre2 | 147 | 78.792 | 0.536 | 58.823 |
| 17. | 080 | MULTICOM-CONSTRUCT | 147 | 76.446 | 0.520 | 58.703 |
| 18. | 253 | pro-sp3-TASSER | 147 | 75.358 | 0.513 | 58.117 |
| 19. | 481 | MUFOLD-Server | 147 | 68.676 | 0.467 | 56.260 |
| 20. | 127 | FAMSD | 147 | 68.669 | 0.467 | 57.295 |

# CASP9 Top 20 Servers on AB Initio Targets

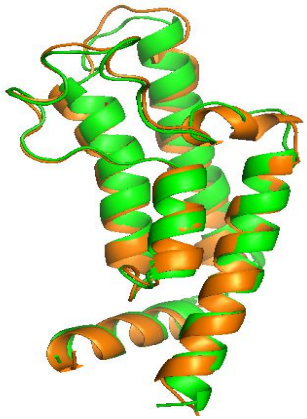
| # | GR # | GR name | Domains Count | SUM Z-score (GDT_TS) | AVG Z-score (GDT_TS) | AVG GDT_TS |
|---|---|---|---|---|---|---|
| 1. | 380 | QUARK | 29 | 31.632 | 1.091 | 31.851 |
| 2. | 428 | Zhang-Server | 29 | 26.506 | 0.914 | 30.539 |
| 3. | 119 | MULTICOM-REFINE | 29 | 22.423 | 0.773 | 28.971 |
| 4. | 457 | chunk-TASSER | 29 | 20.697 | 0.714 | 28.626 |
| 5. | 002 | MULTICOM-CLUSTER | 29 | 19.954 | 0.688 | 28.445 |
| 6. | 286 | RaptorX | 29 | 19.754 | 0.681 | 27.730 |
| 7. | 077 | RaptorX-MSA | 29 | 19.343 | 0.667 | 27.494 |
| 8. | 321 | BAKER-ROSETTASERVER | 29 | 18.973 | 0.654 | 27.118 |
| 9. | 253 | pro-sp3-TASSER | 29 | 18.892 | 0.651 | 27.996 |
| 10. | 276 | RaptorX-Boost | 29 | 18.286 | 0.631 | 27.400 |
| 11. | 215 | MULTICOM-NOVEL | 29 | 18.266 | 0.630 | 27.813 |
| 12. | 055 | MUFOLD-MD | 28 | 16.900 | 0.604 | 24.986 |
| 13. | 080 | MULTICOM-CONSTRUCT | 29 | 16.582 | 0.572 | 27.219 |
| 14. | 063 | Jiang_Assembly | 29 | 14.721 | 0.508 | 26.249 |
| 15. | 236 | gws | 29 | 13.931 | 0.480 | 26.002 |
| 16. | 047 | BioSerf | 29 | 13.590 | 0.469 | 24.716 |
| 17. | 103 | SAM-T08-server | 29 | 12.686 | 0.437 | 24.618 |
| 18. | 452 | Seok-server | 29 | 12.552 | 0.433 | 24.762 |
| 19. | 481 | MUFOLD-Server | 29 | 10.590 | 0.365 | 23.567 |
| 20. | 174 | Phyre2 | 29 | 10.385 | 0.358 | 24.465 |

# Some High-Quality CASP Predictions



**T0390**
**GDT=0.90**

**T0426**
**GDT=0.97**



**T0432**
**GDT=0.92**

**T0458**
**GDT=0.97**

**Orange: structure; Green: model**

**50 of 120 CASP8 targets are in high-accuracy, RMSD < 2 Å**

# Genome-Wide Annotation of Protein Structure


Schmutz et al., Nature, 2010

DATABASE Open Access

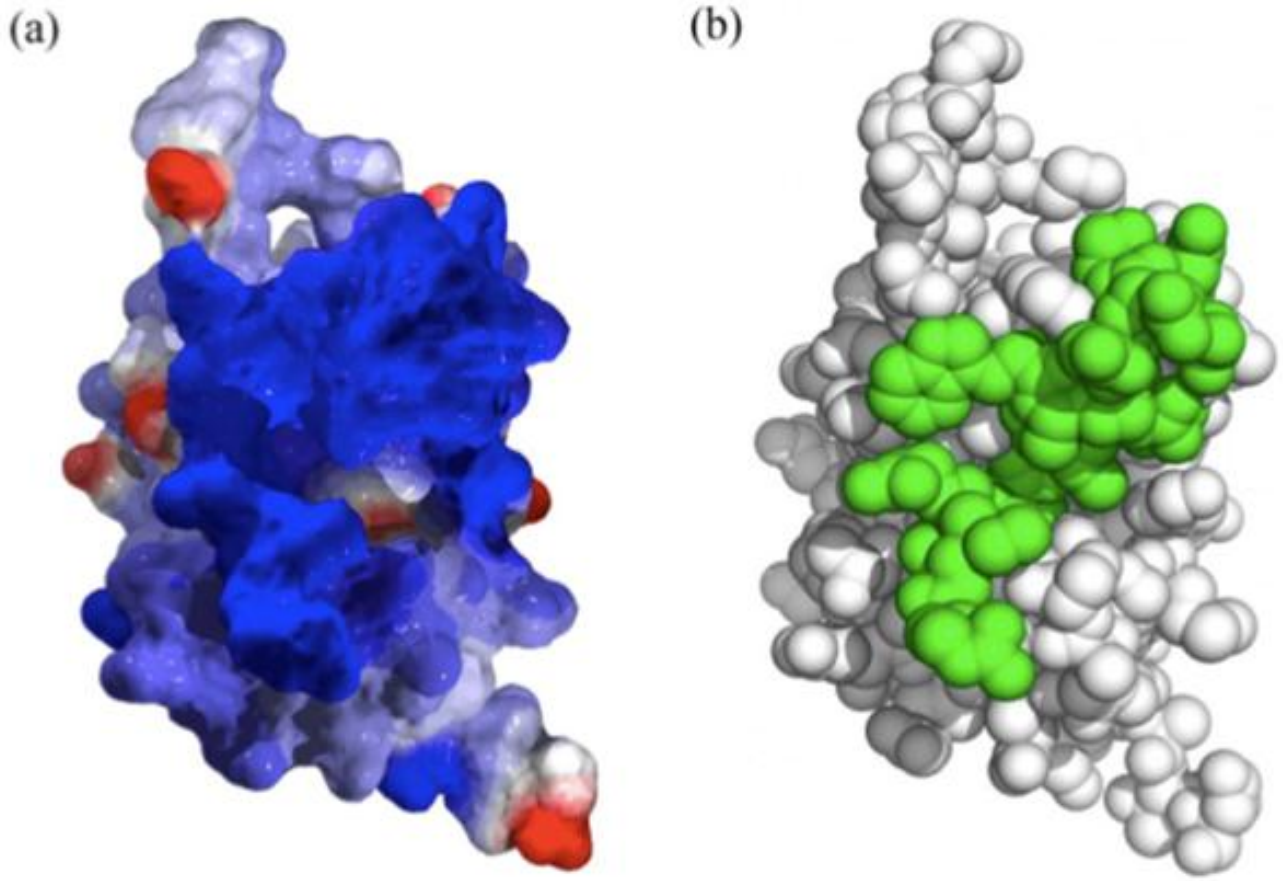
## SoyDB: a knowledge database of soybean transcription factors

**BMC Plant Biology**

Zheng Wang[1], Marc Libault[2,3], Trupti Joshi[1,2], Babu Valliyodan[2,3], Henry T Nguyen[2,3], Dong Xu[1,2,4], Gary Stacey[2,3], Jianlin Cheng[1,2,4*]

Highly accessed

# Transcription Factor Glyma01g32810




Wang et al., BMC Plant Biology, 2010

# Acknowledgements

**Group Members**

- Xin Deng
- Jesse Eickholt
- Jianfeng He
- Jilong Li
- Cuong Nguyen
- Allison Tegge
- Zheng Wang