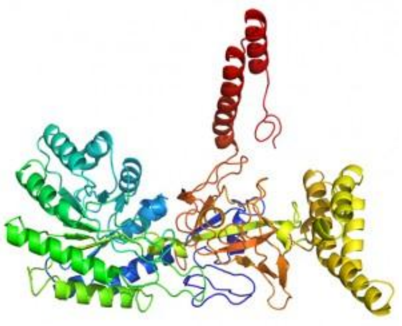
Recursive Protein Structure Modeling



Jianlin Cheng, PhD

Department of Computer Science

Informatics Institute

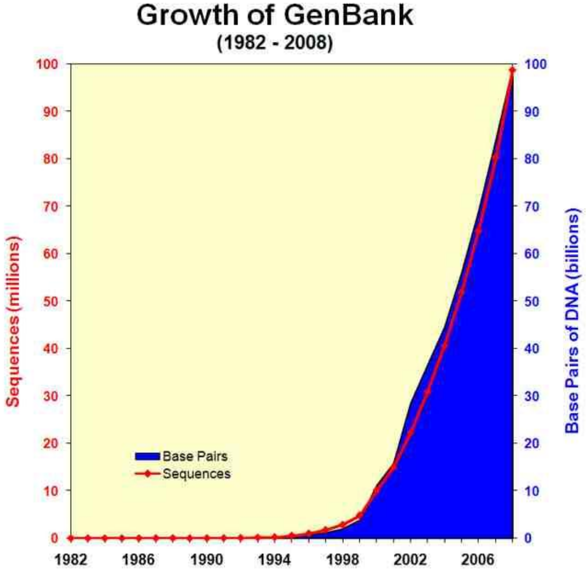
University of Missouri, Columbia

Presented at BIBM Computational Structural Bioinformatics Workshop, Nov. 12, 2011

Protein Structure Prediction – A Key Challenge in the Genomic Era

Genome Sequencing

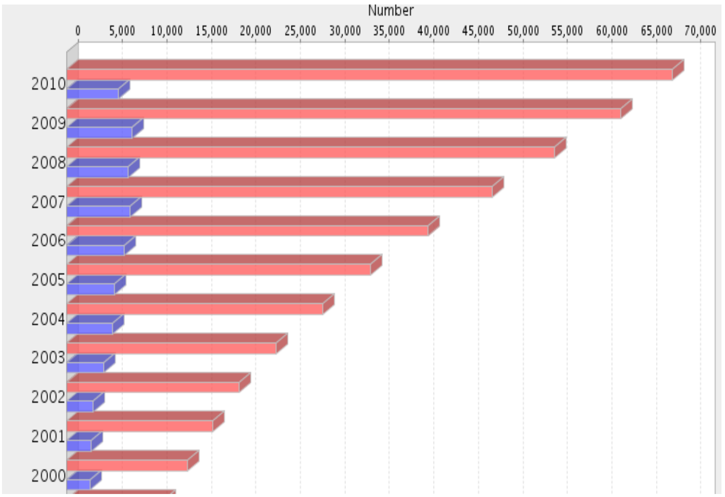
Genome Interpretation



Growth of GenBank

(1982 - 2008)

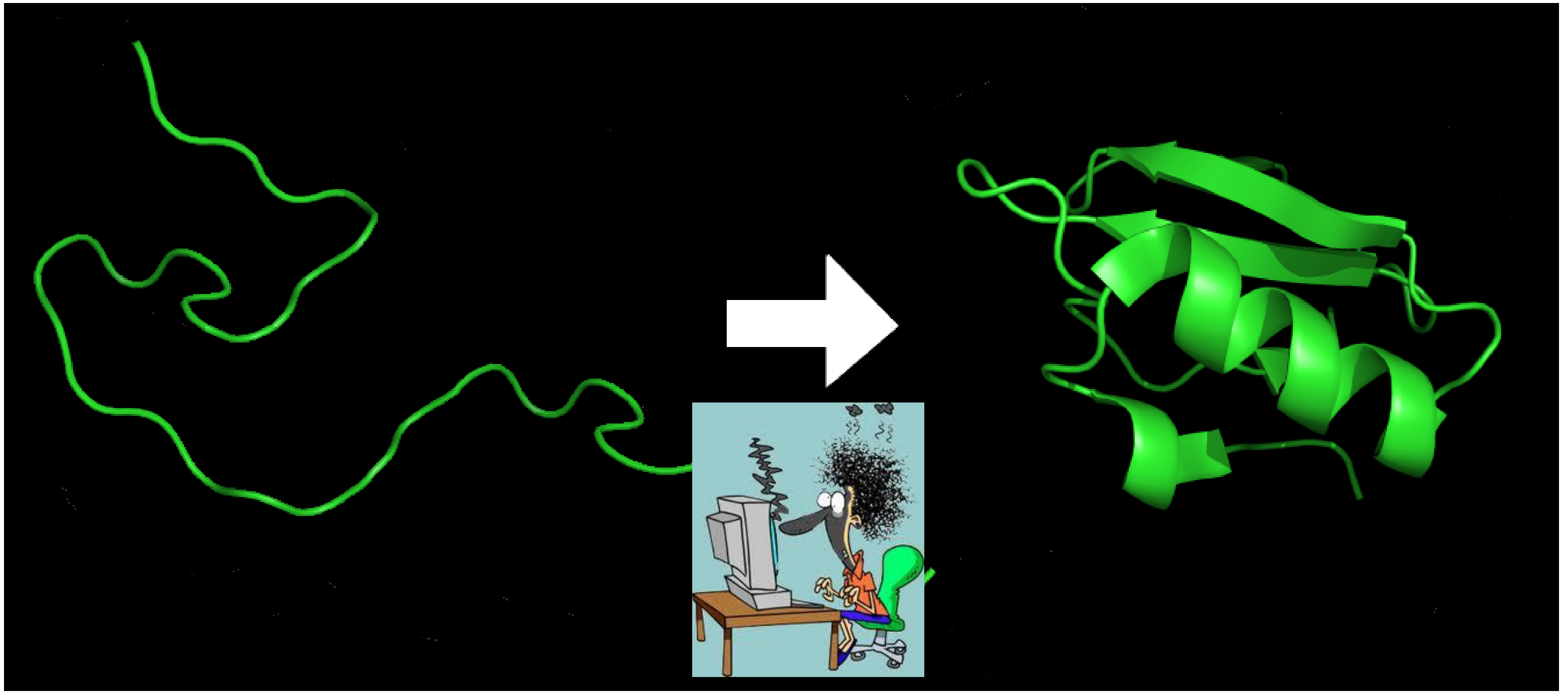
Growth of PDB Structures



Images.google.com

Computational Protein Structure Prediction

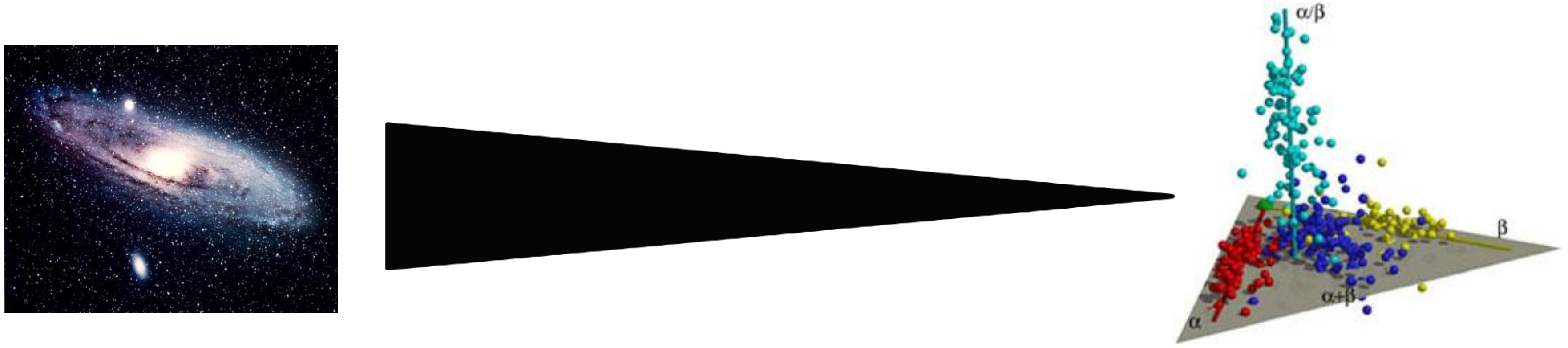
Structure = *f* (sequence) ? ↔ **$E = MC^2$**



Computational Simulation

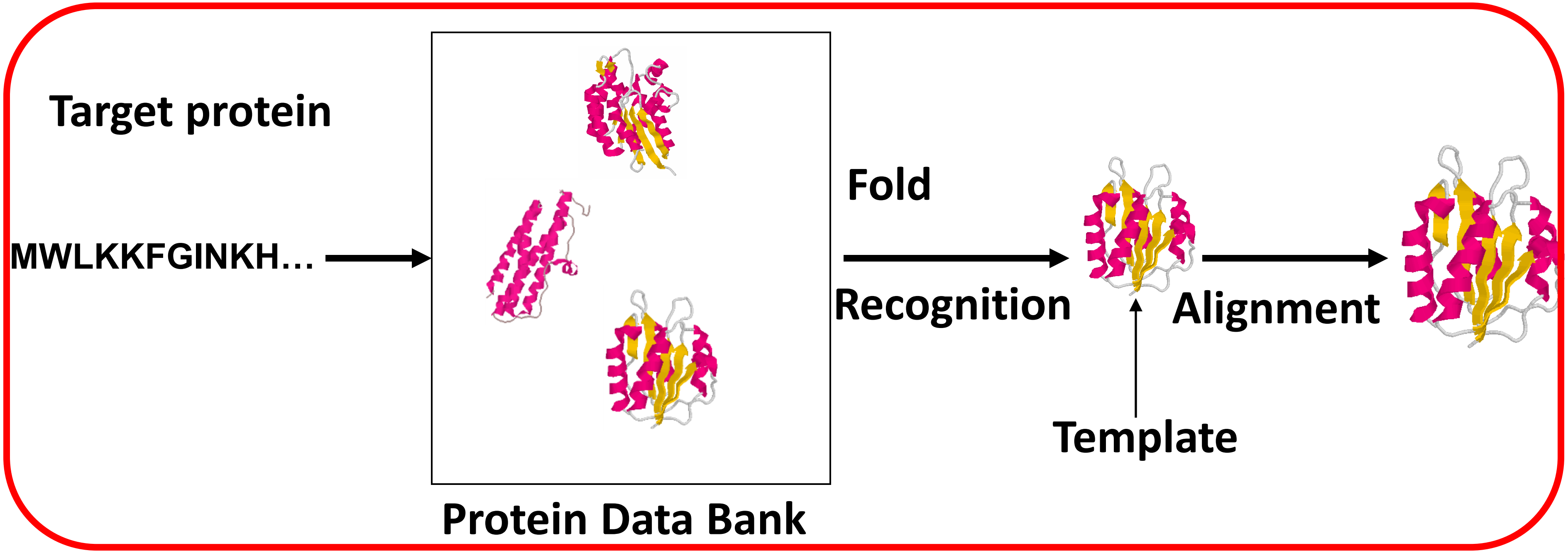
Images.google.com

Template-Based Modeling



Sequence Space

Structure Space



Target protein

MWLKKFGINKH…

Protein Data Bank

Fold

Recognition

Alignment

Template

Images.google.com

Template-Based Modeling

TARGET

ASILPKRLFGNCEQTSDEGLK
IERTPLVPHISAQNVCLKIDD
VPERLIPERASFQWMNDK

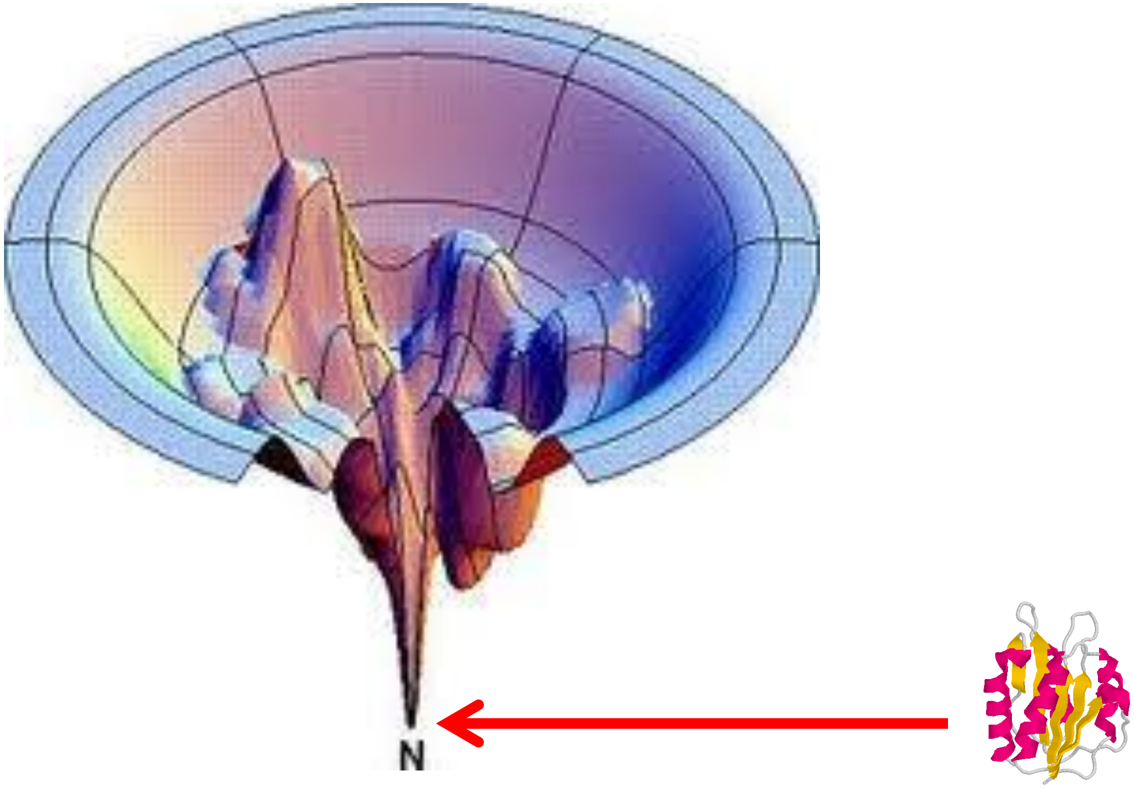
TEMPLATE

```
ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIPE
MSVIPKRLYGNCEQTSEEAIRIEDSPIV---TADLVCLKIDEIPERLVGE
```

Modeller

A. Fisher, 2005

Template-Free (*Ab Initio*) Modeling



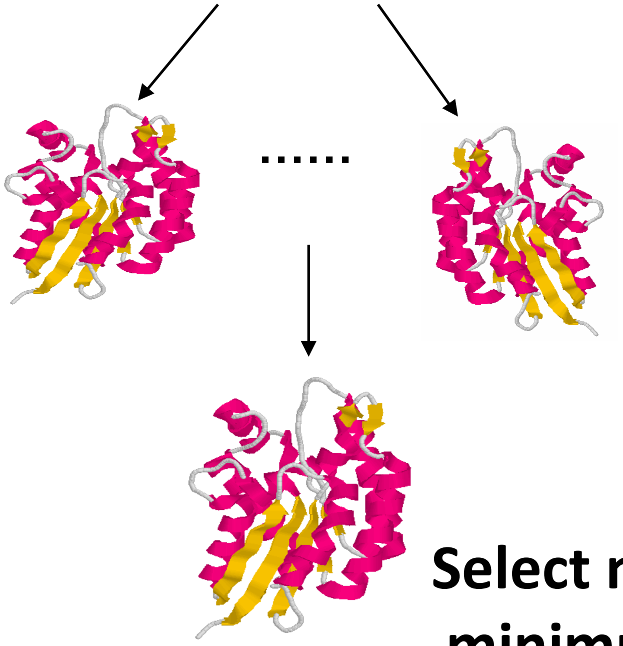
Native Structure

Protein Structure Space

Dill & Chan, 1997

Template-Free Modeling

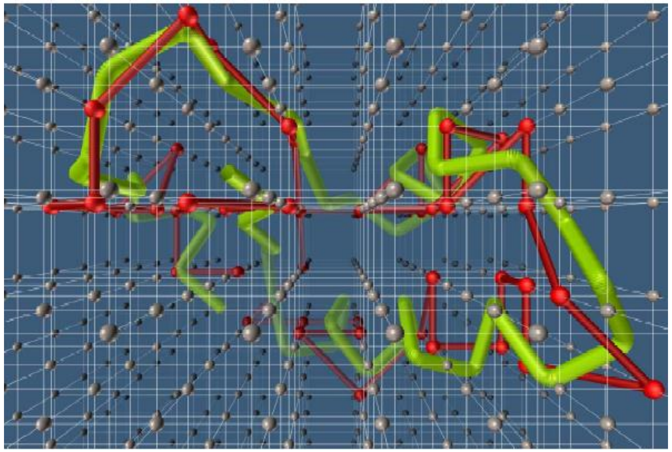
MWLKKFGINLLIGQSV…



......

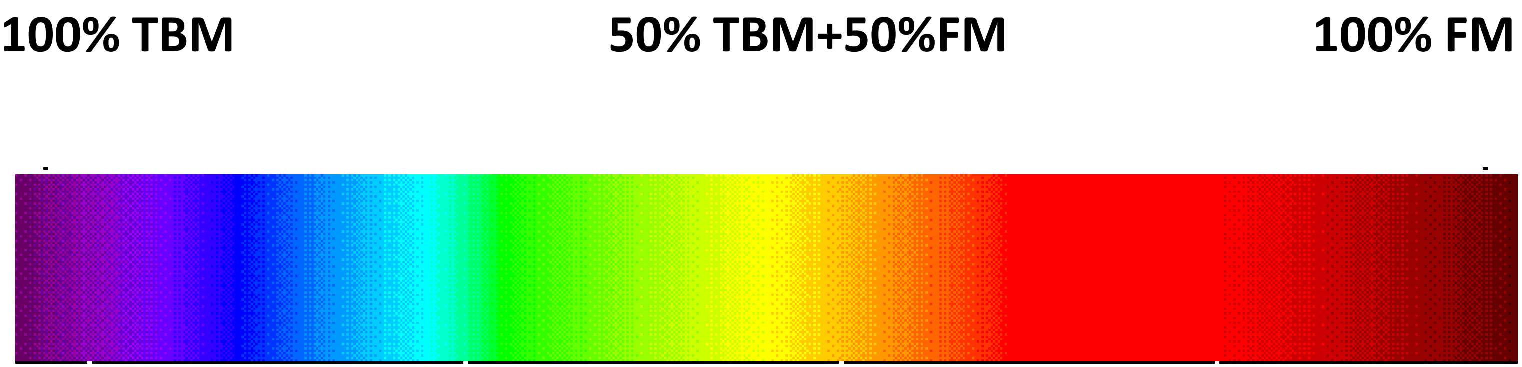
Select model with minimum free energy

3D Simulation



Methods: molecular dynamics, fragment assembly, distance / contact-based modeling

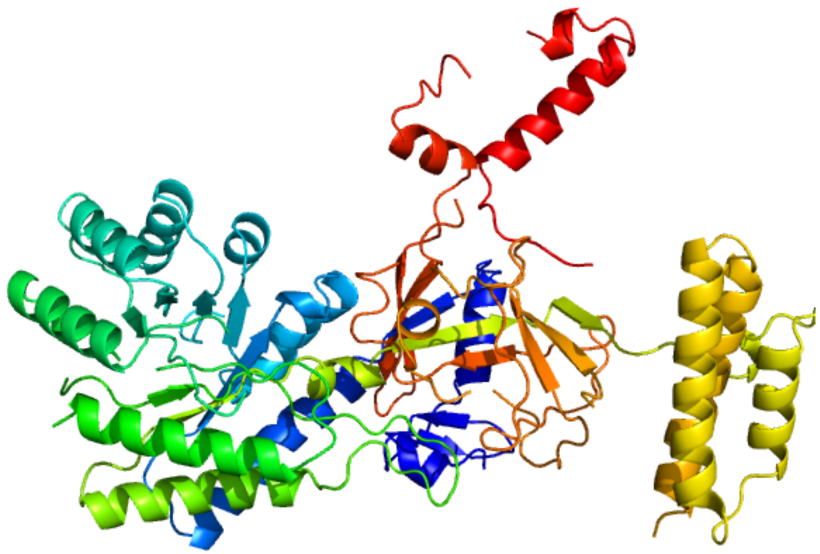
Combination of Template-Free and Template-Based Modeling



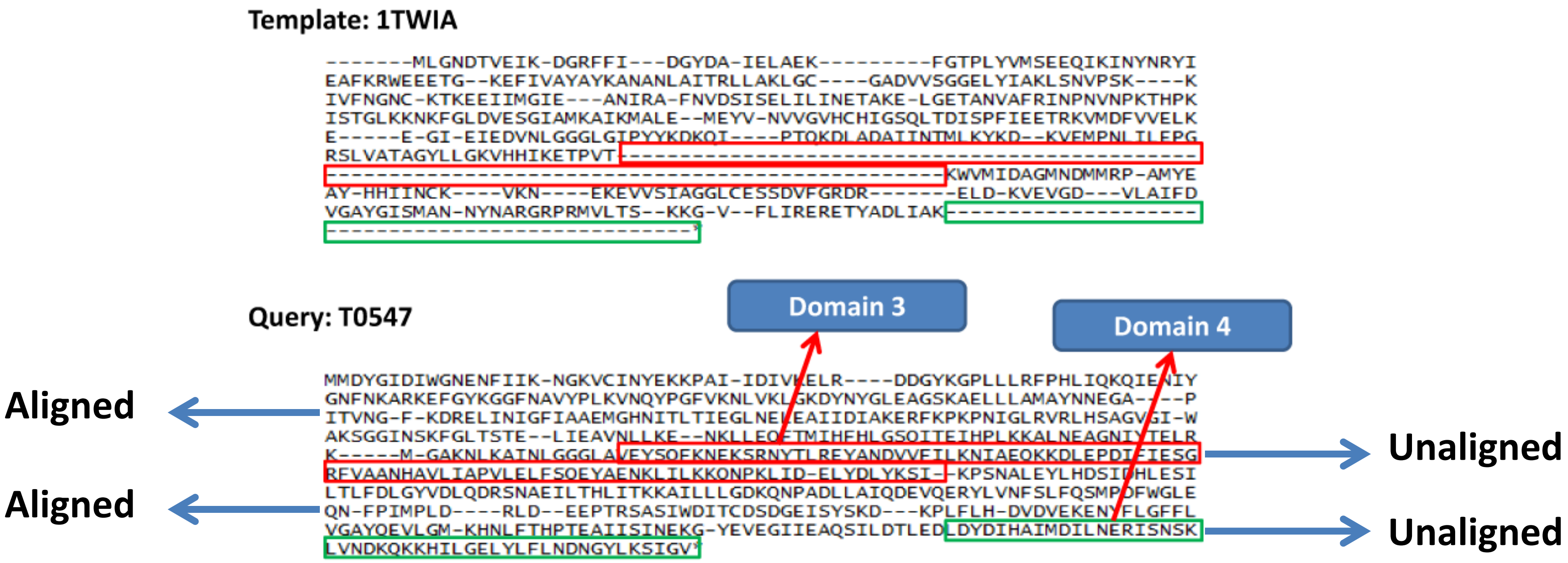
Protein Modeling Spectrum

Region Decomposition from Alignment

A)



B)



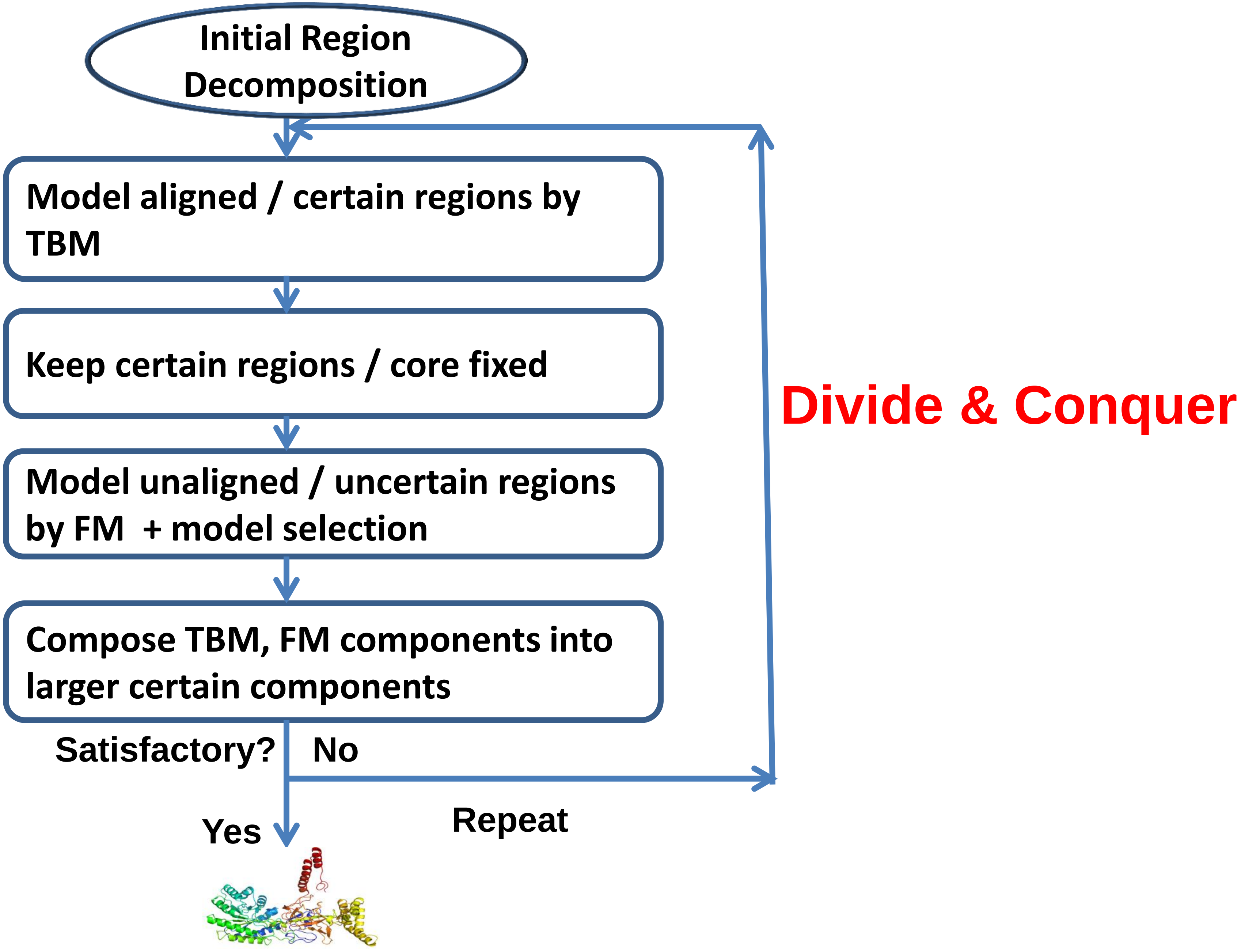
Template: 1TWIA

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

Query: T0547

```
MMDYGIDIWGNENFIIK-NGKVCINYEKKPAI-IDIVKELR----DDGYKGPLLLRFPHLIQKQIENIY
GNFNKARKEFGYKGGFNAVYPLKVNQYPGFVKNLVKLGKDYNYGLEAGSKAELLLAMAYNNEGA----P
ITVNG-F-KDRELINIGFIAAEMGHNITLTIEGLNELEAIIDIAKERFKPKPNIGLRVRLHSAGVGI-W
AKSGGINSKFGLTSTE--LIEAVNLLKE--NKLLEQFTMIHEHLGSQITEIHPLKKALNEAGNIYTELR
K-----M-GAKNLKAINLGGGLAVEYSQFKNEKSRNYTLREYANDVVFILKNIAEQKKDLEPDIFIESG
REVAANHAVLIAPVLELFSQEYAENKLILKKQNPKLID-ELYDLYKSI--KPSNALEYLHDSIDHLESI
LTLFDLGYVDLQDRSNAEILTHLITKKAILLLGDKQNPADLLAIQDEVQERYLVNFSLFQSMPDFWGLE
QN-FPIMPLD----RLD--EEPTRSASIWDITCDSDGEISYSKD---KPLFLH-DVDVEKENYFLGFFL
VGAYQEVLGM-KHNLFTHPTEAIISINEKG-YEVEGIIEAQSILDTLEDLDYDIHAIMDILNERISNSK
LVNDKQKKHILGELYLFLNDNGYLKSIGV*
```

Recursive Protein Modeling – Integrate TBM and FM



Initial Region Decomposition

Model aligned / certain regions by TBM

Keep certain regions / core fixed

Model unaligned / uncertain regions by FM + model selection

Compose TBM, FM components into larger certain components

Satisfactory?

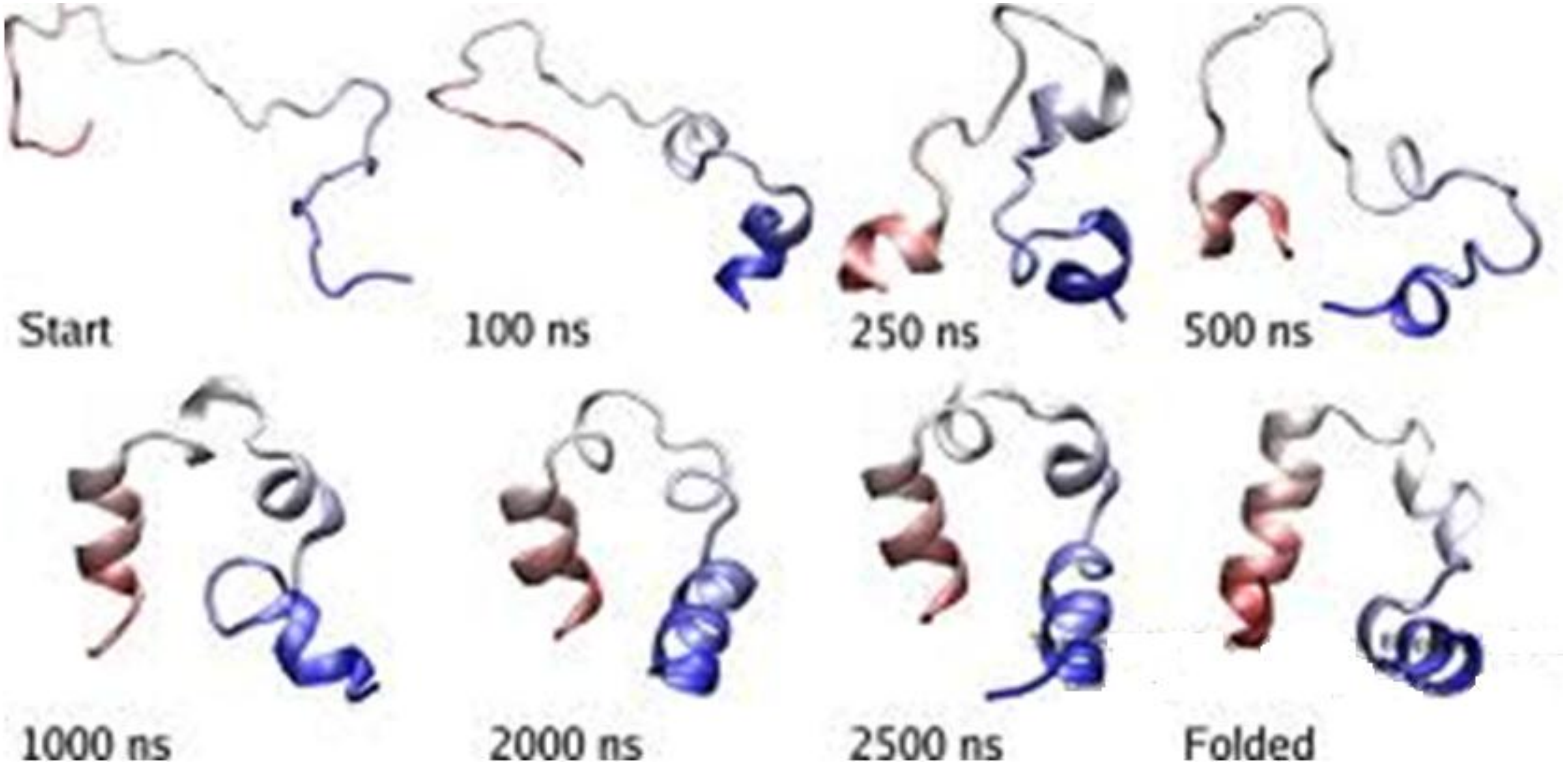
No

Repeat

Yes

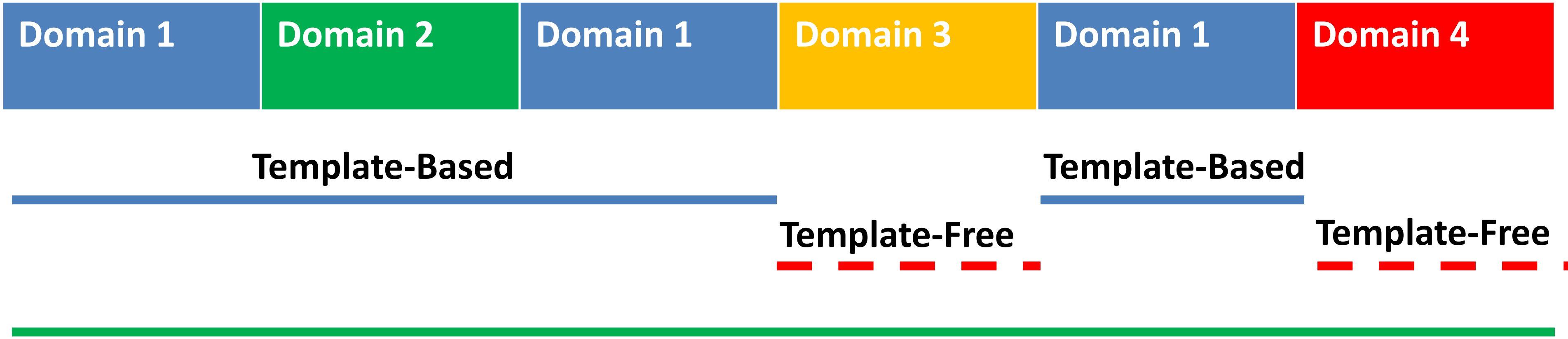
Divide & Conquer

Recursive Modeling Mimics Protein Folding Cascade



ks.uiuc.edu

Case 1: Domain-Level Recursive Protein Modeling – CASP9 T0547



Domain 1

Domain 2

Domain 1

Domain 3

Domain 1

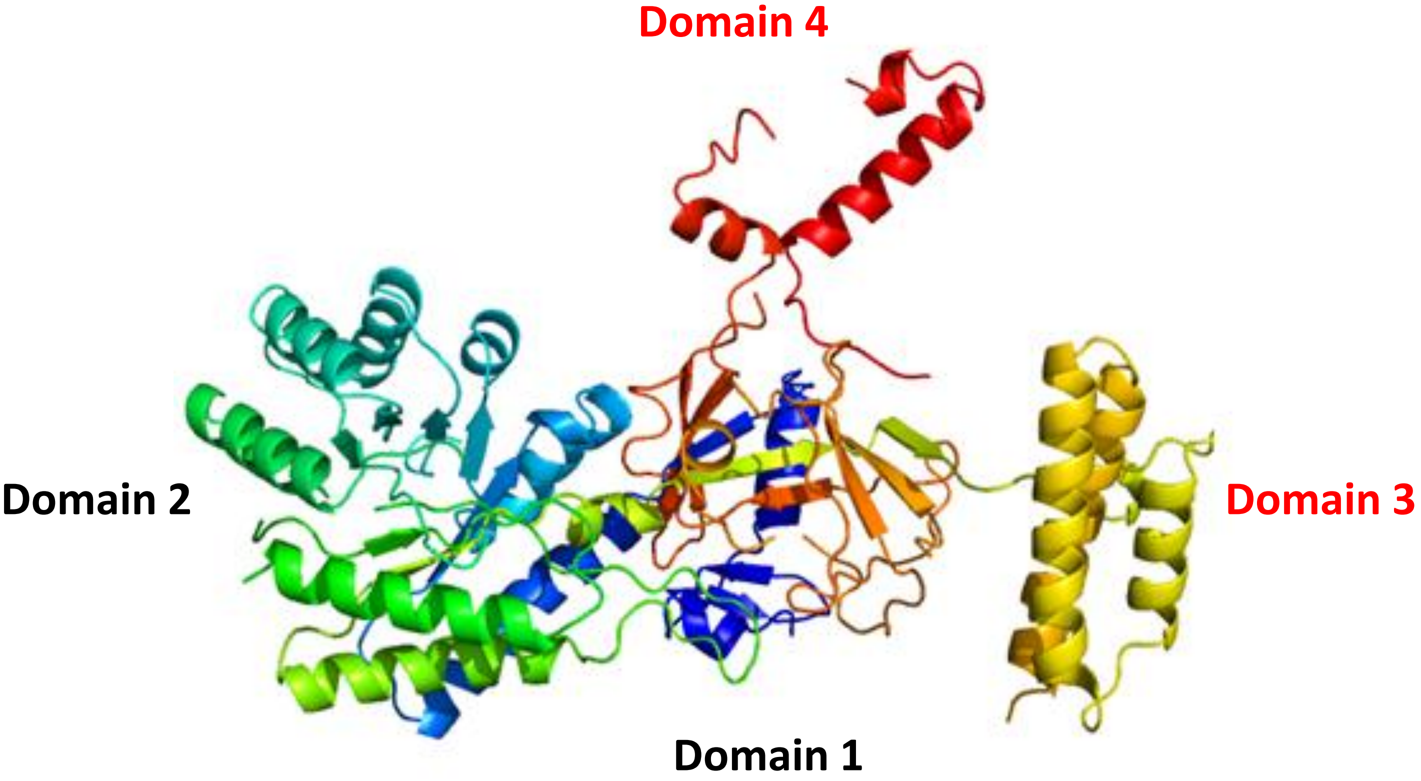
Domain 4

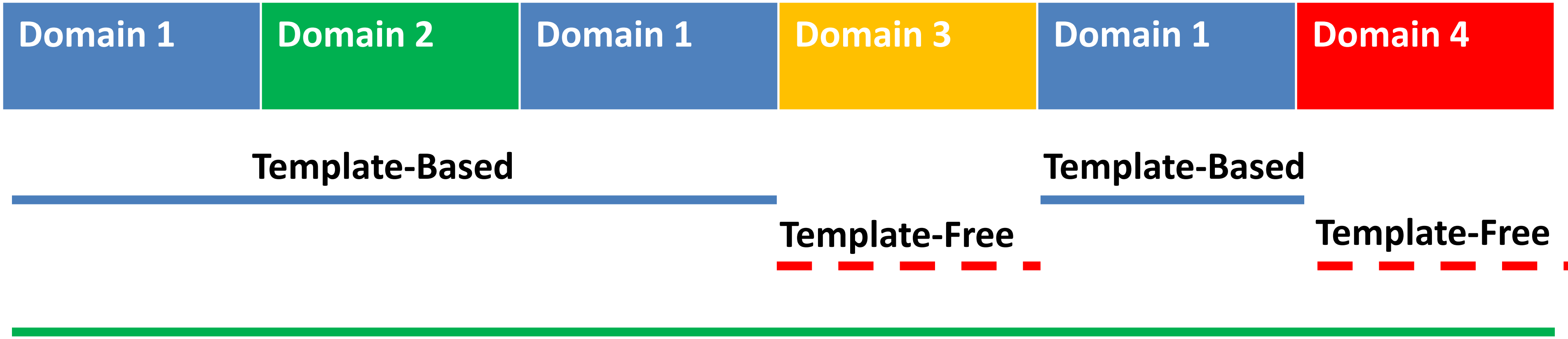
Template-Based

Template-Based

Template-Free

Template-Free





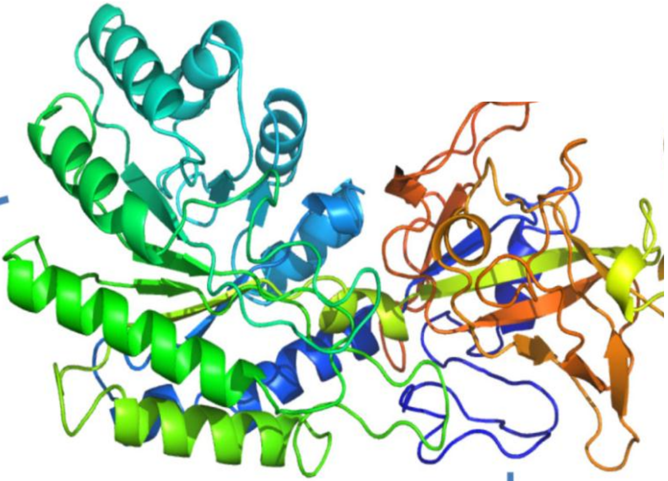
Domain 1 | Domain 2 | Domain 1 | Domain 3 | Domain 1 | Domain 4

Template-Based

Template-Free

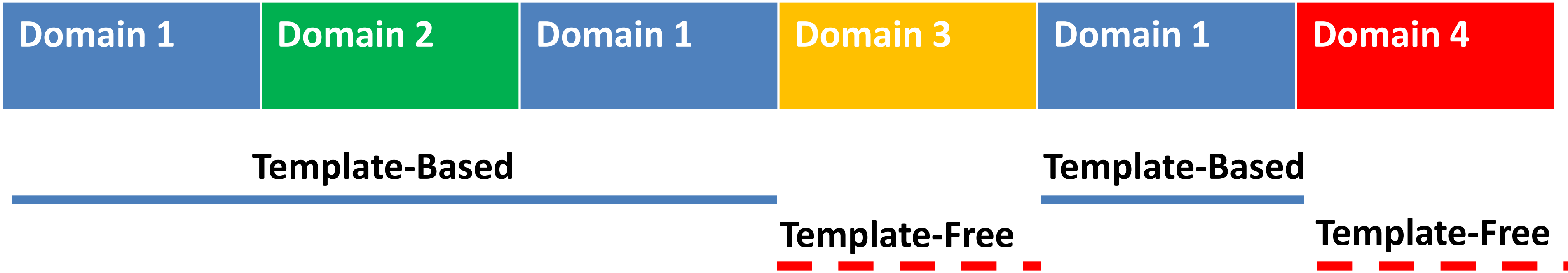
Template-Based

Template-Free



TBM Domain 2
(Insertion)
GDT-TS = 0.74

TBM Domain 1 – Three Discontinuous Fragments
GDT-TS = 0.66



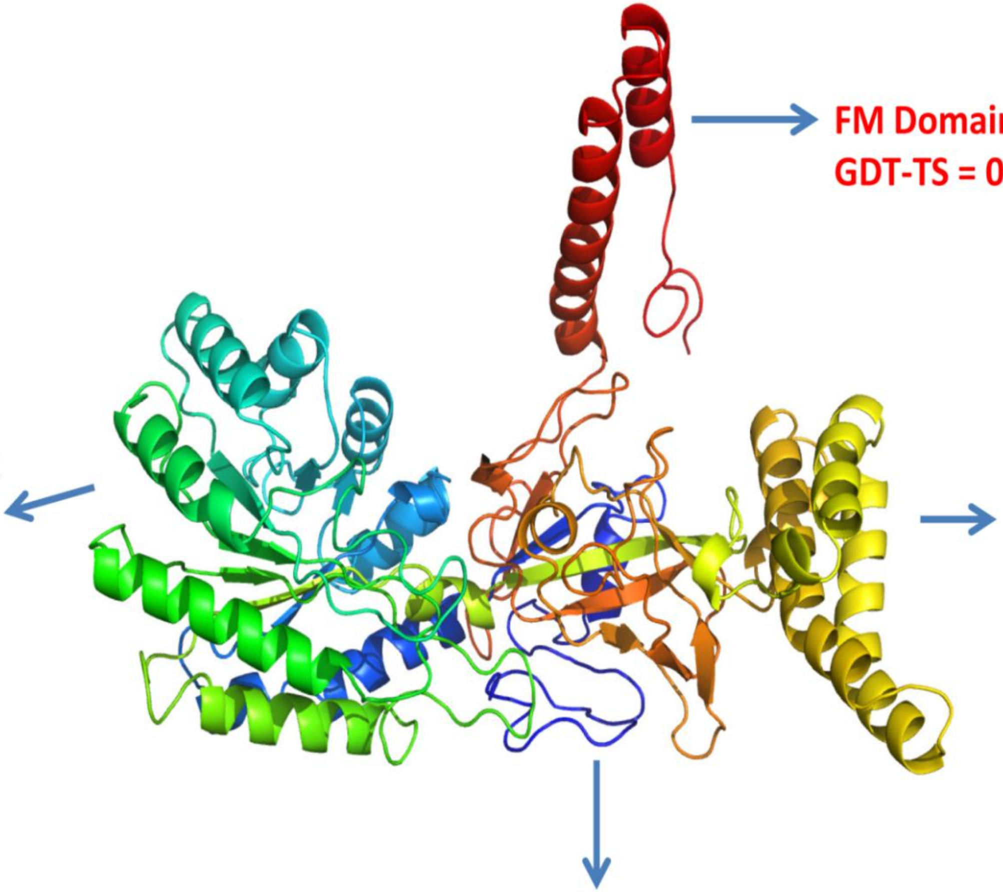
Domain 1 | Domain 2 | Domain 1 | Domain 3 | Domain 1 | Domain 4

Template-Based

Template-Free

Template-Based

Template-Free



TBM Domain 2
(Insertion)
GDT-TS = 0.74

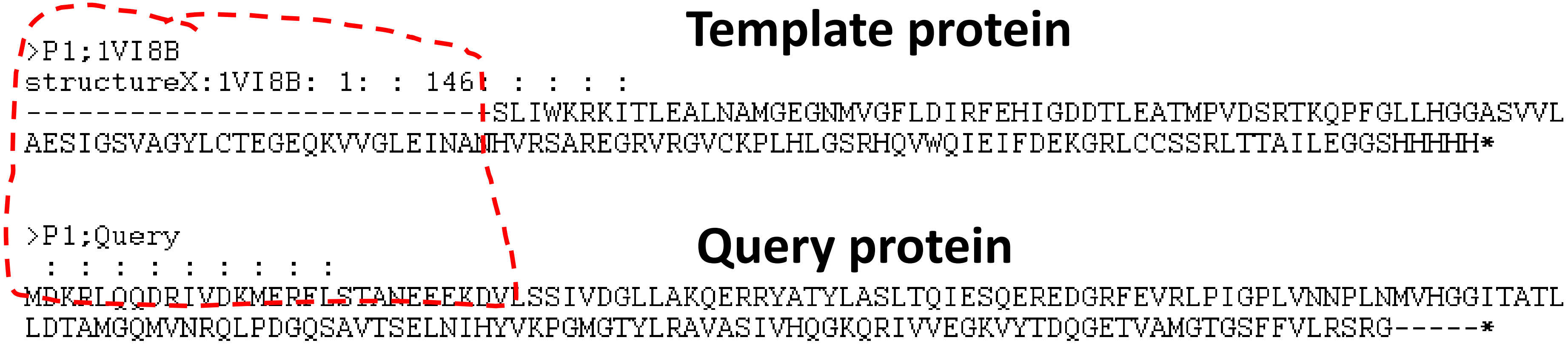
FM Domain 4
GDT-TS = 0.61

FM Domain 3
(Insertion)
GDT-TS = 0.58

TBM Domain 1 – Three Discontinuous Fragments
GDT-TS = 0.66

Case 2: Refine uncertain regions of a largely template-based modeling (T0539)

Core-Constrained Tail Refinement

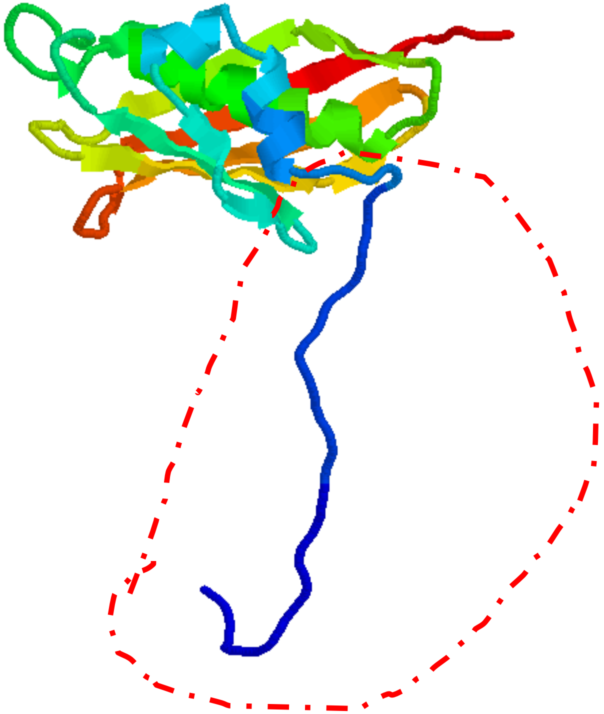


Template protein

Query protein

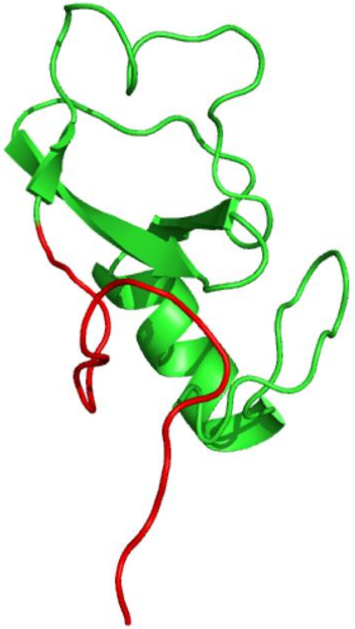
```
>P1;1VI8B
structureX:1VI8B: 1: : 146: : : : :
---------------------------SLIWKRKITLEALNAMGEGNMVGFLDIRFEHIGDDTLEATMPVDSRTKQPFGLLHGGASVVL
AESIGSVAGYLCTEGEQKVVGLEINANHVRSAREGRVRGVCKPLHLGSRHQVWQIEIFDEKGRLCCSSRLTTAILEGGSHHHHH*


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



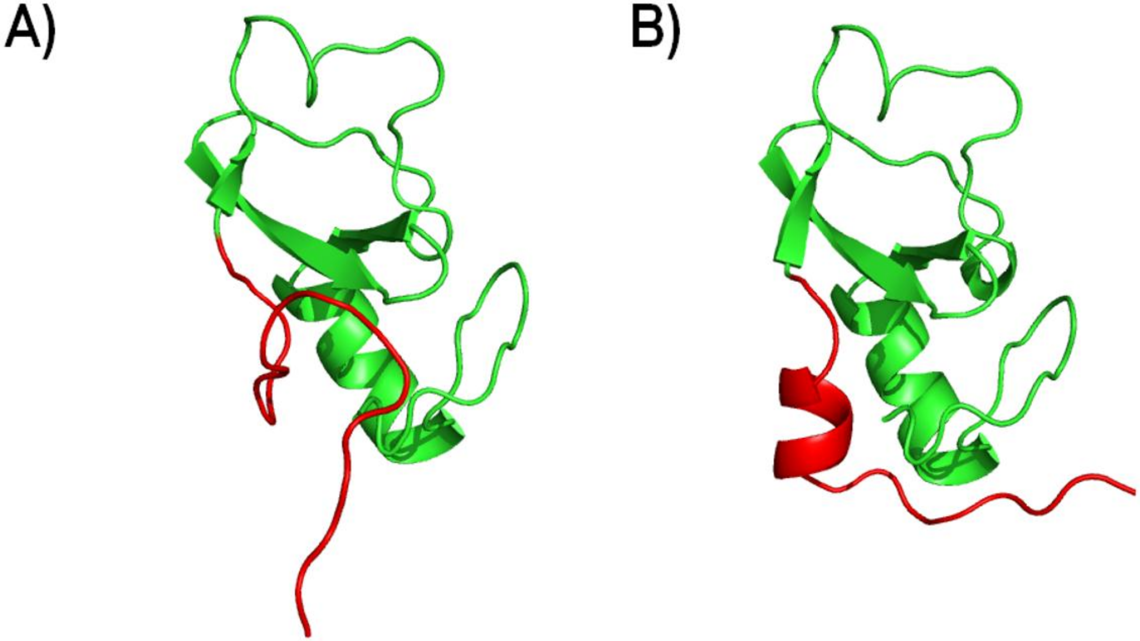
Core-Constrained Tail Refinement

A)



Before tail refinement
GDT-TS = 0.64

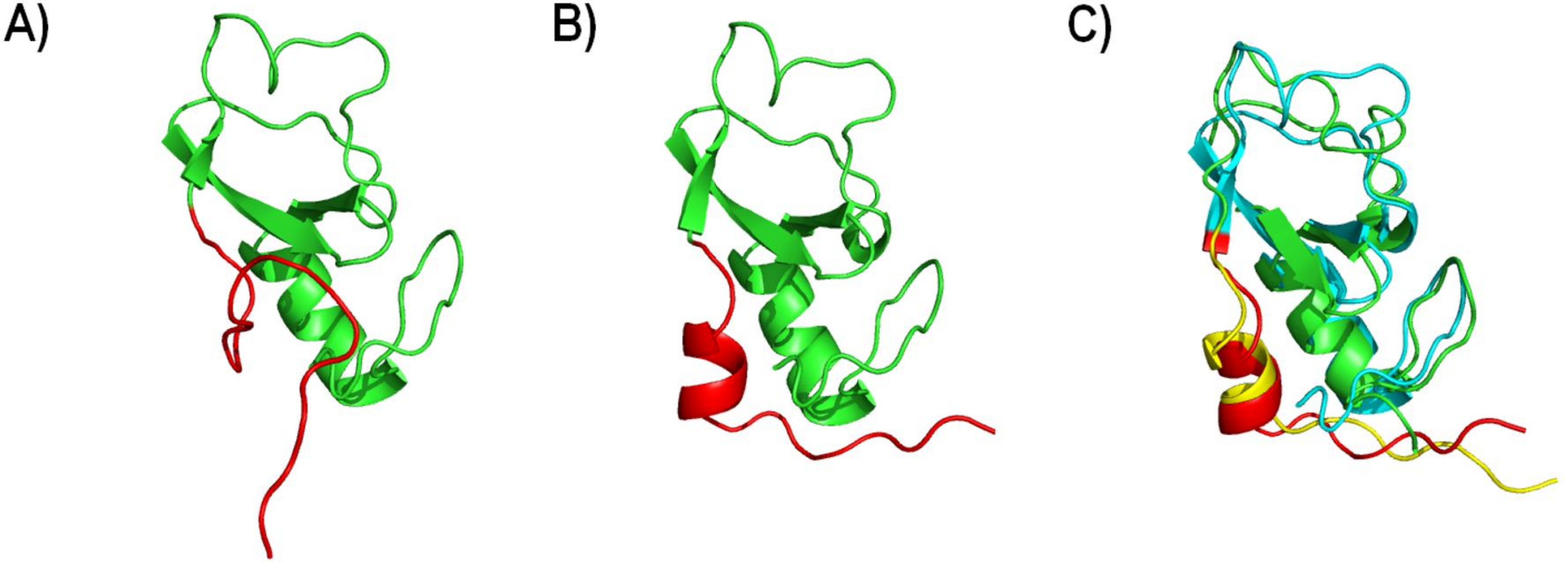
Core-Constrained Tail Refinement



Before tail refinement
GDT-TS = 0.64

After tail refinement
GDT-TS = 0.73

Core-Constrained Tail Refinement

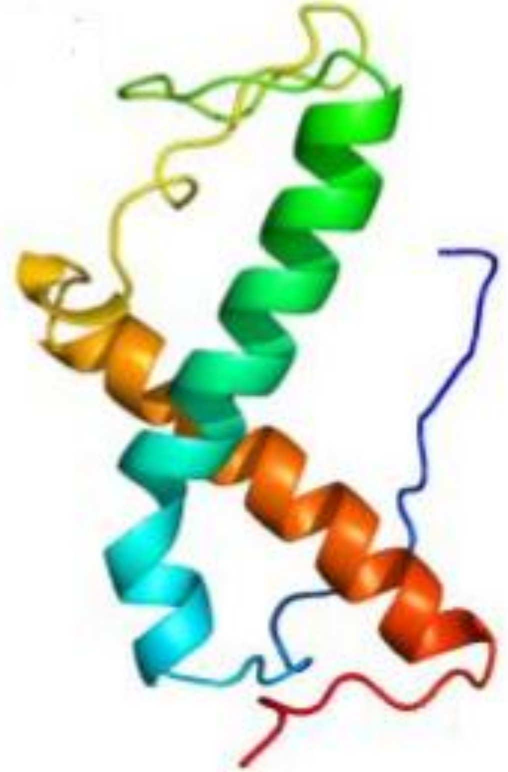


Before tail refinement
GDT-TS = 0.64

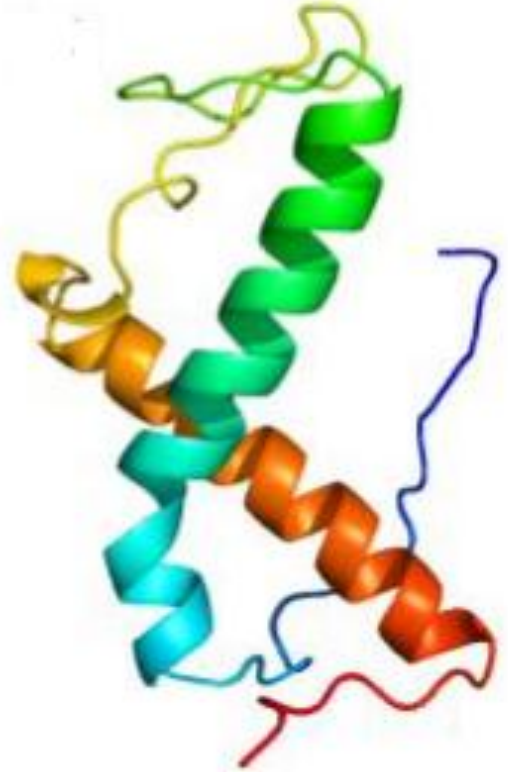
After tail refinement
GDT-TS = 0.73

Superposition
Green: model, Blue: structure

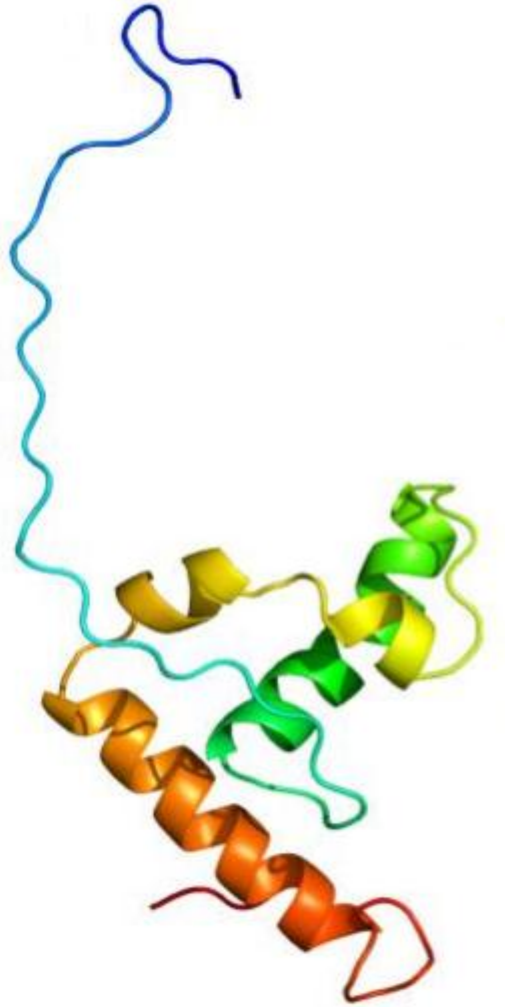
Case 3: Expanding a template-based core into a full structure (T0616)



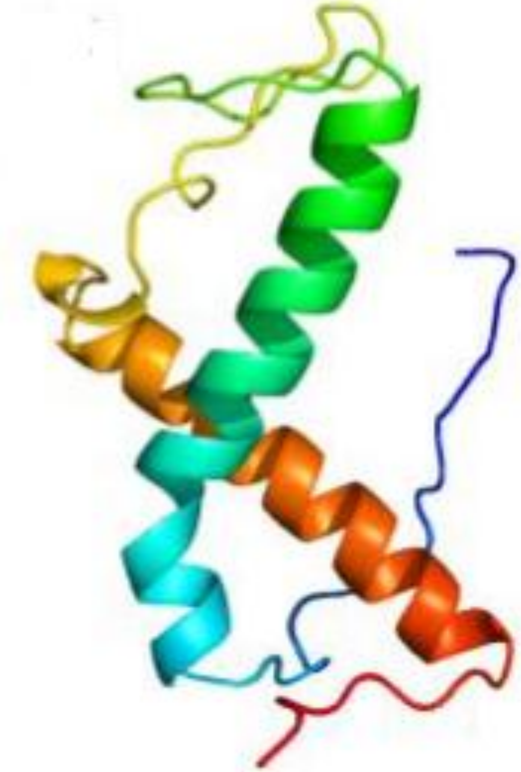
Native Structure



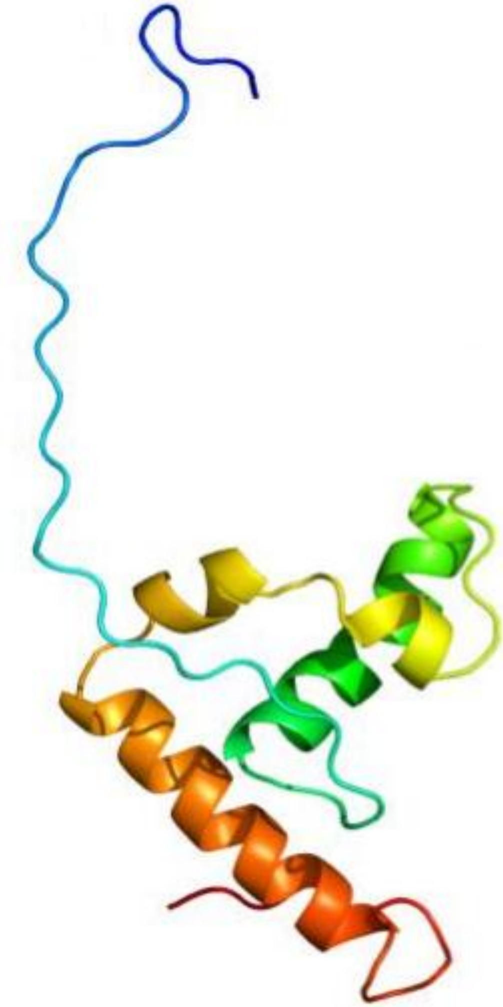
Native Structure



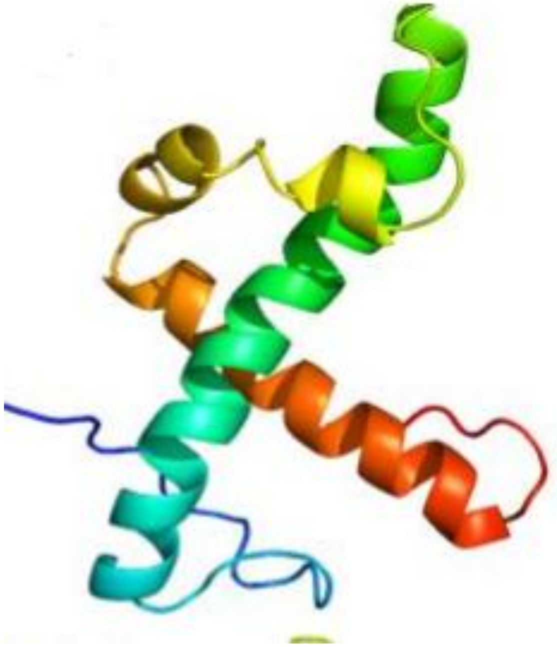
Template-based modeling (GDT-TS = 0.34)



Native Structure



Template-based modeling (GDT-TS = 0.34)



Template-based + Ab Initio (GDT-TS = 0.39)

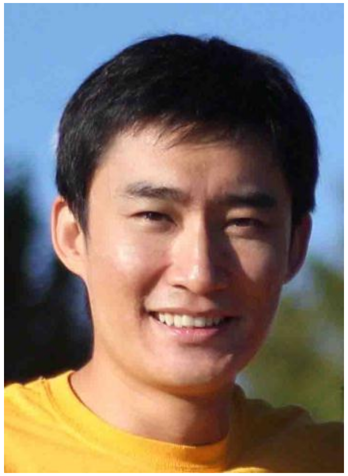
Advantages of Recursive Protein Modeling

- Avoiding error-prone hard decisions on the classification of a protein target or a region
- Combining the strength of template-based modeling and template-free modeling
- Improving sampling efficiency by recursively expanding certain regions
- Easy to implement and improve

Acknowledgements

Xin Deng

Jesse Eickholt



Zheng Wang

NATIONAL INSTITUTES OF HEALTH

MISSOURI

Comparison with Previous Approaches

- **Compared with loop modeling**

  Region of any size: loop, partial domain,
  domain, multiple domains

  Region of any type: helix, strand, loop

- **Compared with TASSER**

  Common: template-based + template-free

  Different: gap filling VS. alternated,
  recursive certainty expansion