Zhang Groups: Automated Structure Prediction by C-I-TASSER and C-QUARK in CASP13

Robin Pearce, Wei Zheng, Chengxin Zhang, Yang Li, S M Mortuza, Yang Zhang

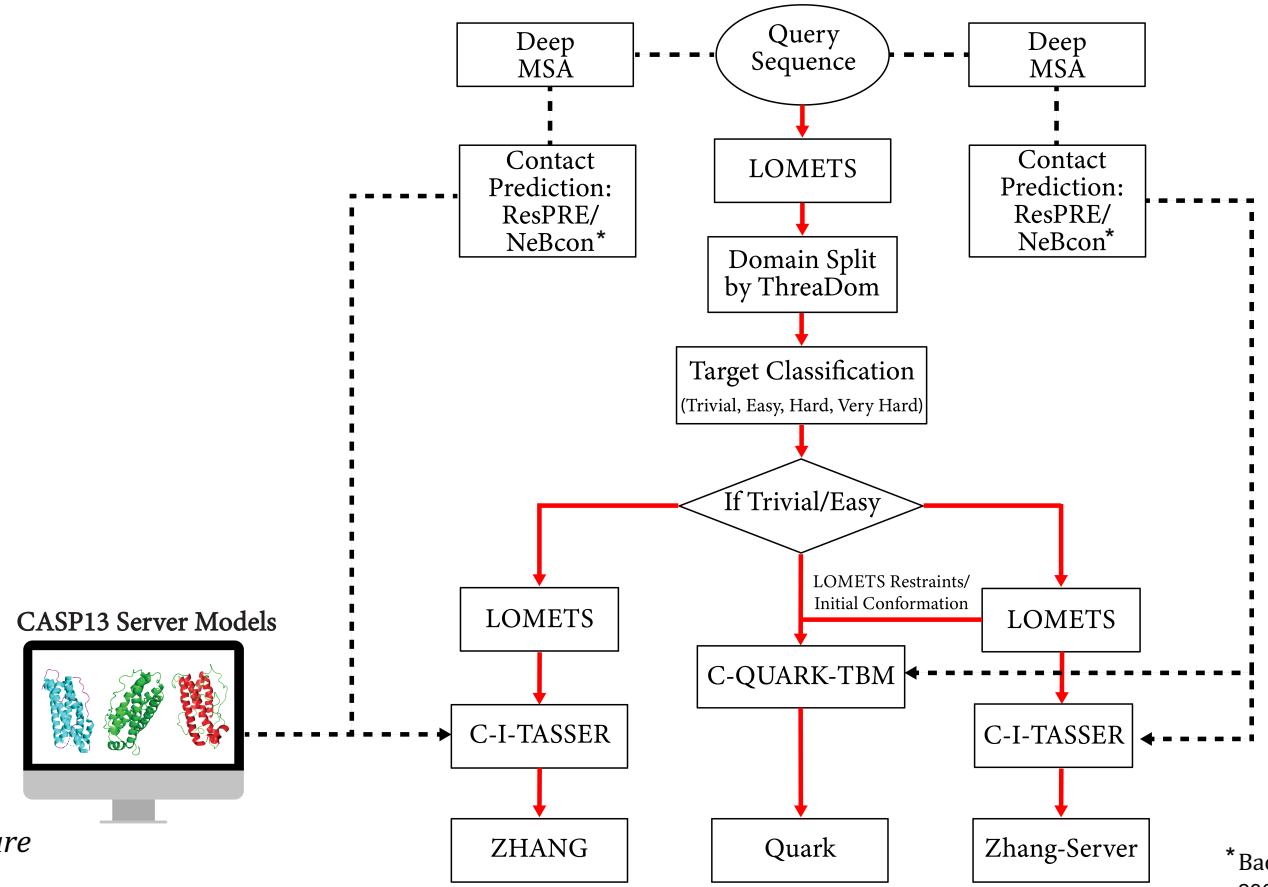
Department of Computational Medicine and Bioinformatics
Department of Biological Chemistry
University of Michigan

Outline

- **Methods**
  - CASP13 Workflow (Trivial, Easy, Hard, Very Hard Targets)
  - Contact Prediction Method and Potential
  - C-Quark and Zhang-Server Pipelines
- **Results**
  - Summary of FM and TBM Modeling Results
  - Impact of Contact Prediction and Template Quality on Modeling Results
  - Case Studies
  - Problems that We Encountered
- **Summary**

Methods

Pipeline for CASP13 Trivial/Easy Targets

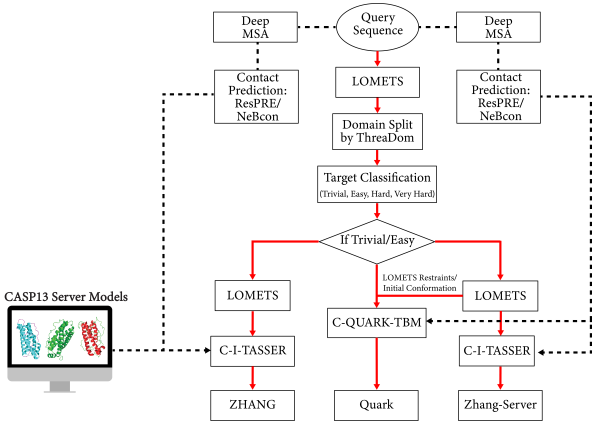


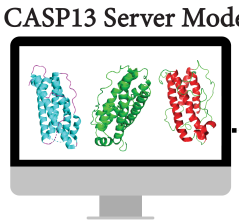
All Predictions are Automated

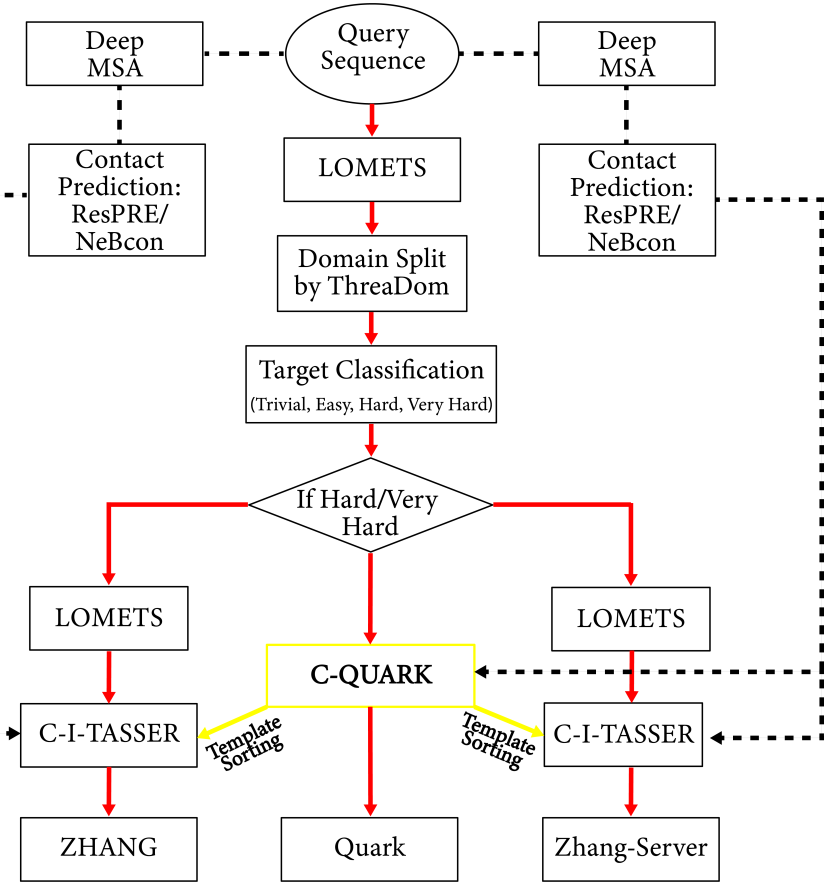
* Baoji He et al. **Bioinformatics**, 33: 2296 2306 (2017)

Pipeline for CASP13 Hard/Very Hard Targets

Pipeline for CASP13 Trivial/Easy Targets

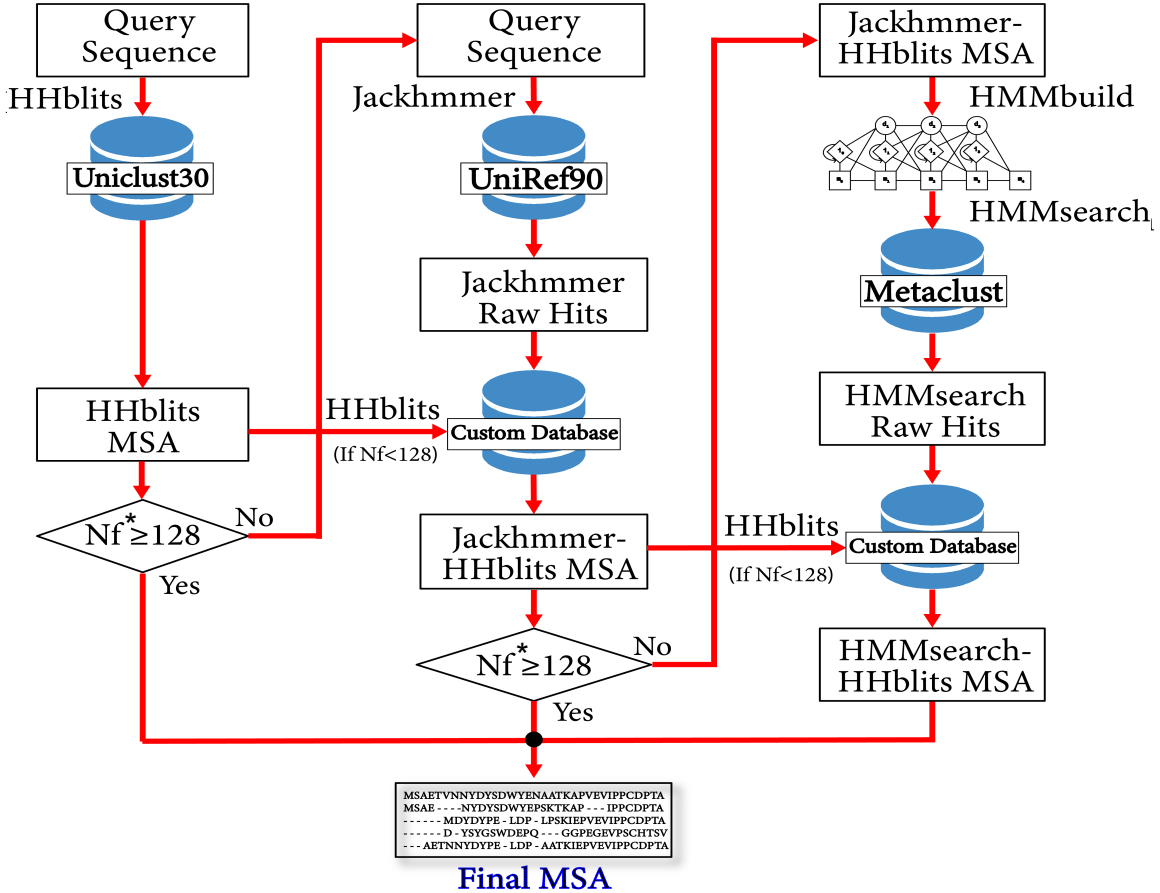






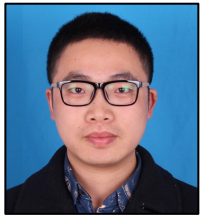
All Predictions are Automated

Deep MSA Construction

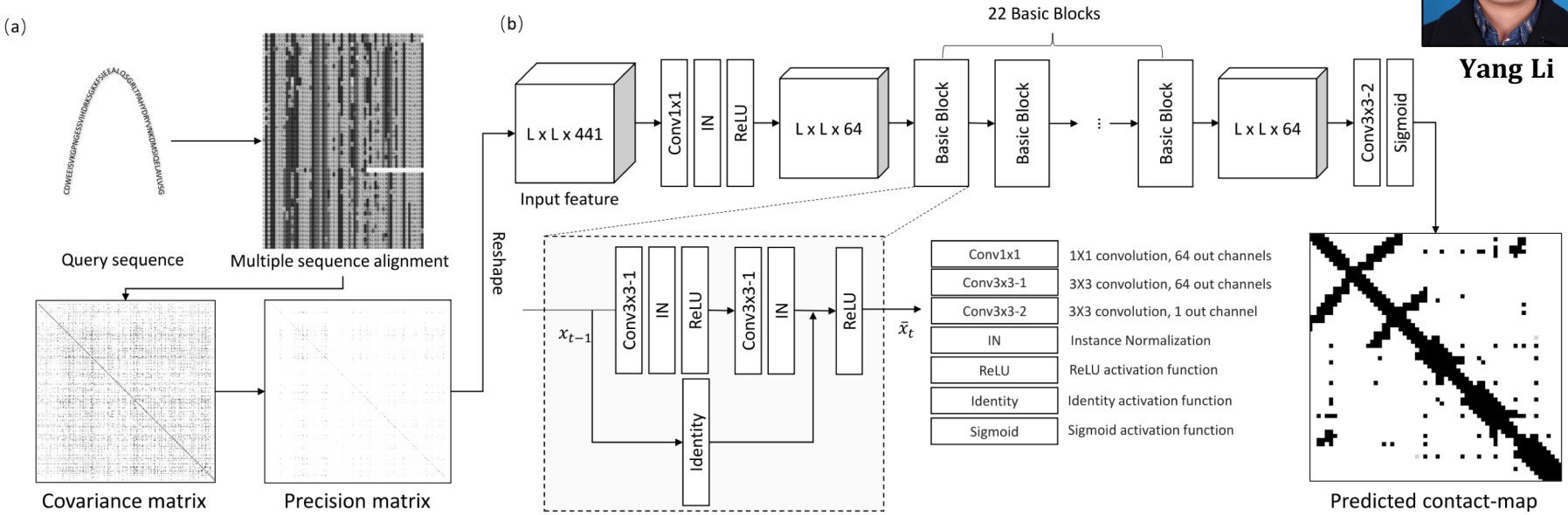


* *Nf: Number of Effective Sequences in MSA*

Contact Prediction Using ResPRE



Yang Li



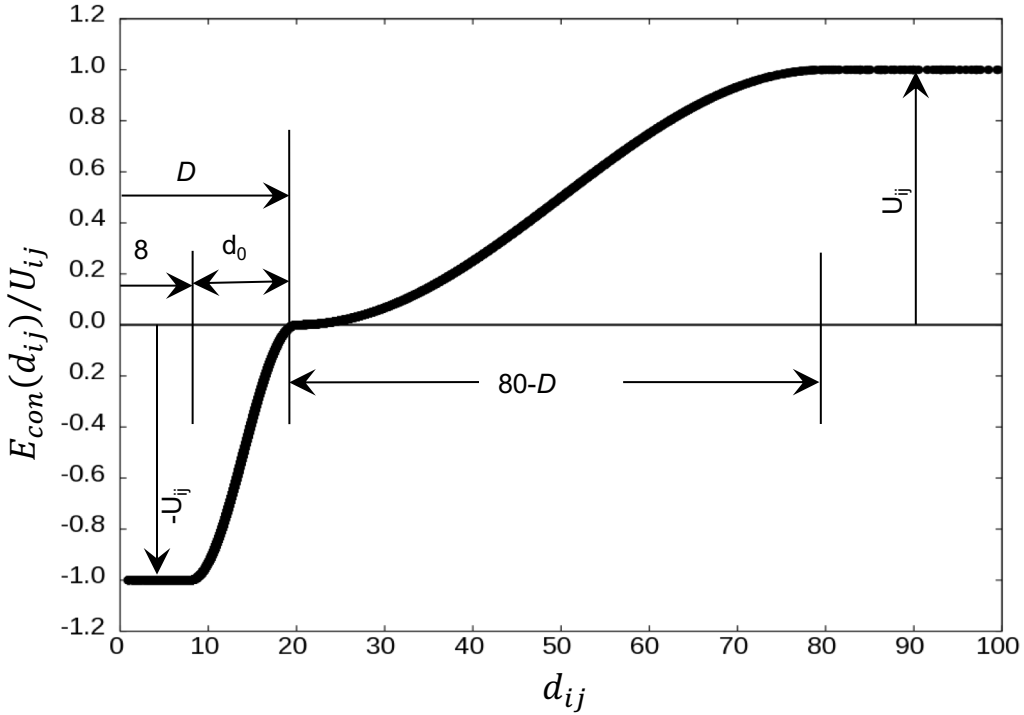
Covariance Matrix (S):

$$S_{i,j}(a,b) = f_{i,j}(a,b) \quad f_i(a) \cdot f_j(b)$$

Precision Matrix (θ):

$$\theta = \underset{\theta}{\operatorname{argmin}} \; (tr(S \cdot \theta) \quad \log(\det(\theta)) + \rho \cdot \|\theta\|_2^2)$$

Three Gradient Contact Potential in C-I-TASSER/C-QUARK

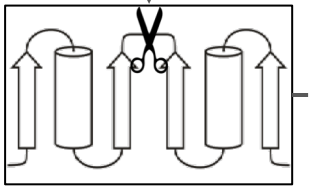


$$E_{con}(d_{ij}) = \begin{cases} -U_{ij}, & d_{ij} < 8\text{Å} \\ -\frac{1}{2}U_{ij}\left[1 - sin\left(\frac{d_{ij} - (\frac{8+D}{2})}{d_b}\pi\right)\right], & 8\text{Å} \le d_{ij} < D \\ \frac{1}{2}U_{ij}\left[1 + sin\left(\frac{d_{ij} - (\frac{D+80}{2})}{(80-D)}\pi\right)\right], & D \le d_{ij} \le 80\text{Å} \\ U_{ij}, & d_{ij} > 80\text{Å} \end{cases}$$

C-QUARK Pipeline with LOMETS

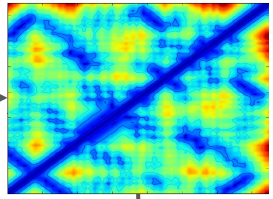
MPRGRSIGRSIGRDTEKERSRSPLRAPGGSGPS
TRAGCGAGPCQLSSPIKRGGVRGQPAKCHGKYT
TTAEGLTALLNRRHSPRTSNEGRWMNGVMAVNL
SKWDTEKERSRSPLRAPGGSDGPSTRAGCGAGP
CQLSSPIKRGGVRGQPAKCHGKYTTTAEGLTAL

Query Sequence

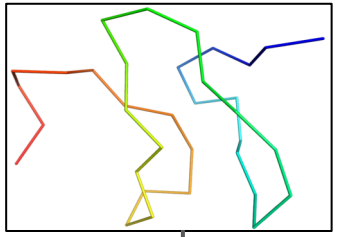


ThreaDom Domain Partition

Gapless Threading Through PDB



Contact Map: ResPRE/ NeBcon

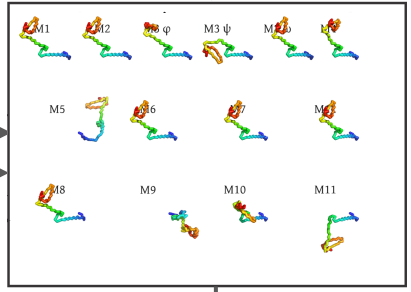


LOMETS Templates (Restraints/ Starting Conformation)

Trivial/ Easy

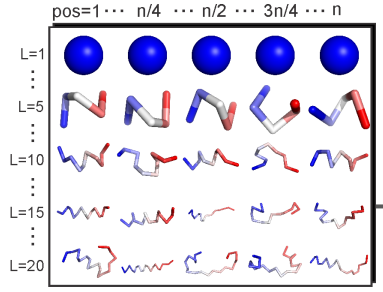
- Contact potential
- Excluded volume
- Backbone hydrogen bond
- Backbone dihedral angle
- RW atomic pairwise interaction
- Generic side-chain contact
- Sequence specific solvation
- Radius of gyration
- Left-hand βαβ motif penalty
- Helix packing
- Cα-Cα pseudo bond
- Distance profile restraints

Physics/Knowledge-based Potential

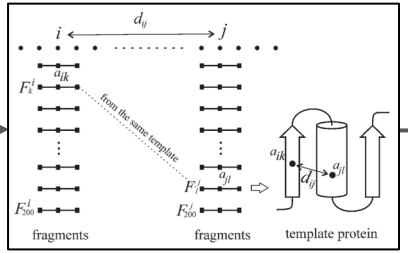


REMC Simulation

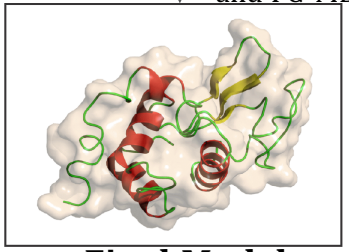
SPICKER Clustering and FG-MD Refinement



200 Fragments with 1-20 AA Created for Each Position

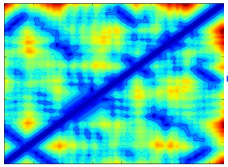


Distance Profile Restraints

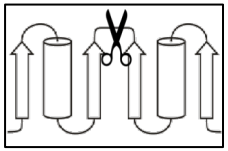


Final Model

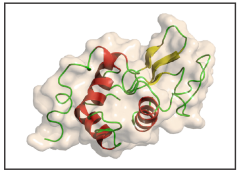
Zhang-Server Pipeline Built on C-I-TASSER



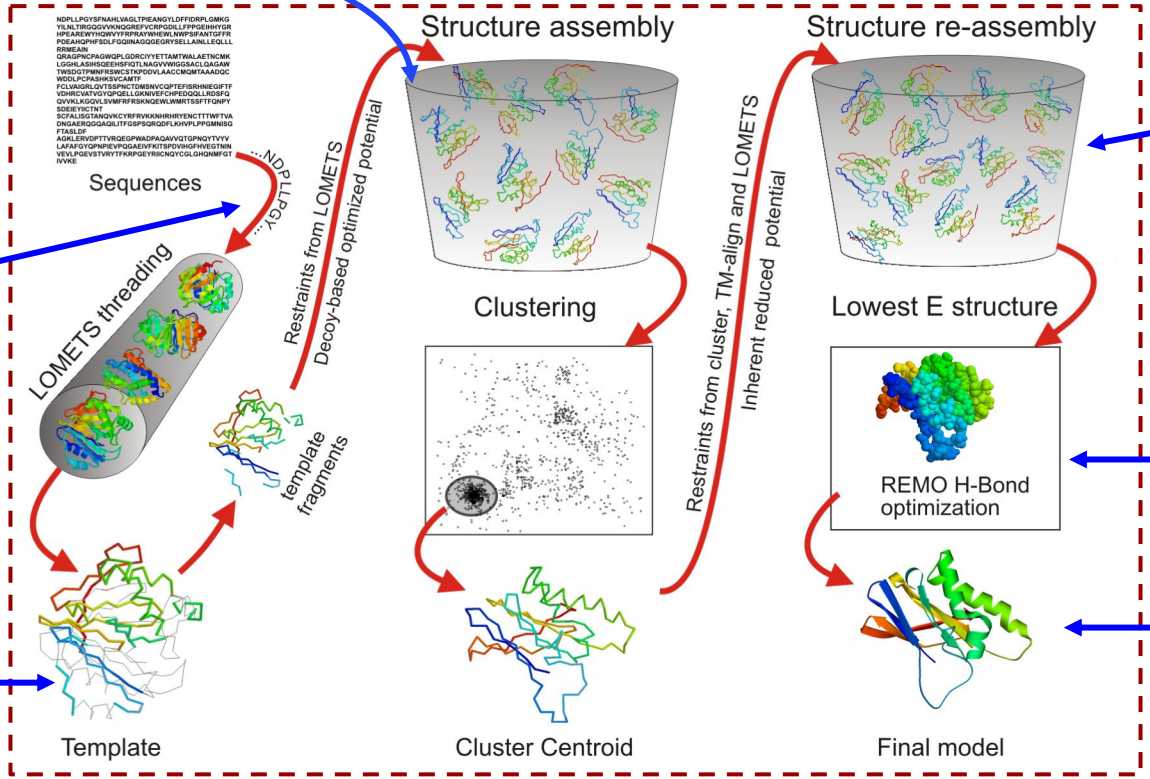
Contact Map:
ResPRE/NeBcon

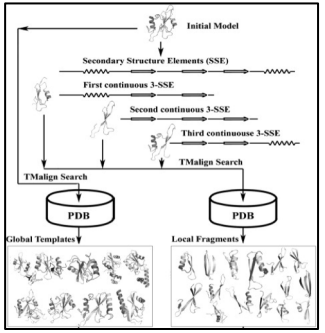


Domain Partition
by ThreaDom

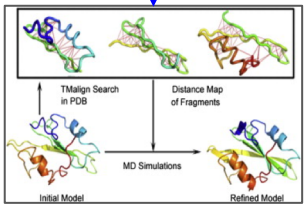


Template Sorting
by C-QUARK
(Hard/Very Hard)





Fragments by TM-align

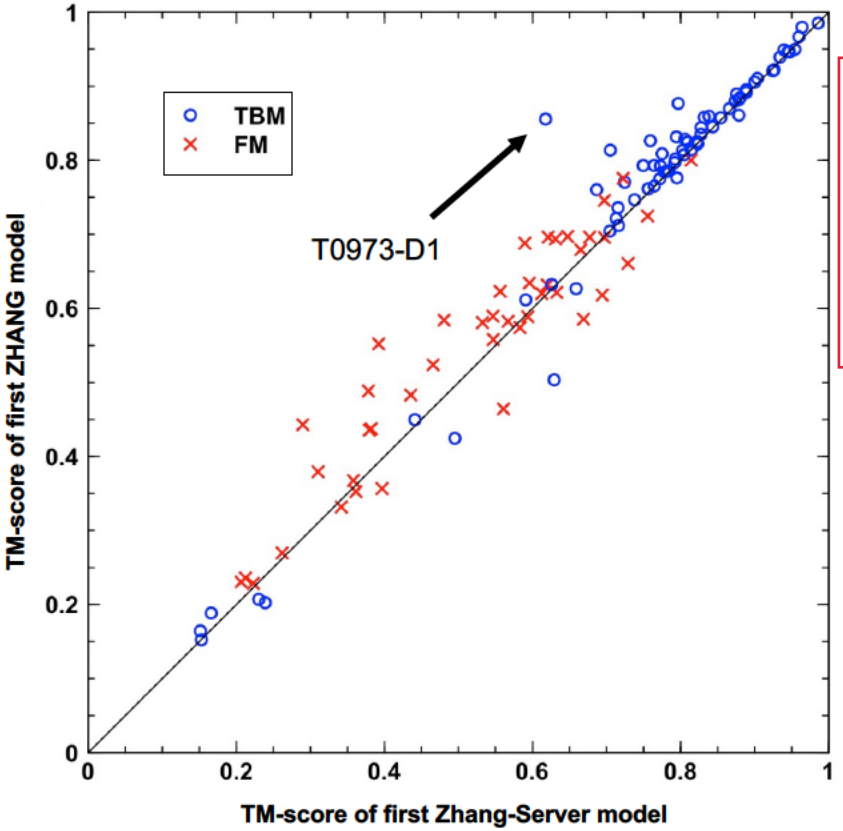


Atomic Refinement
by FG-MD

Residue Quality
Estimation by ResQ

Results

ZHANG vs. Zhang-Server



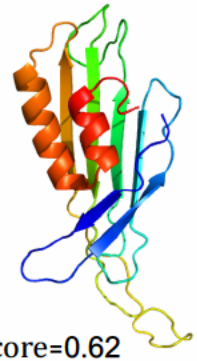
Experimental

CASP:
Image redacted

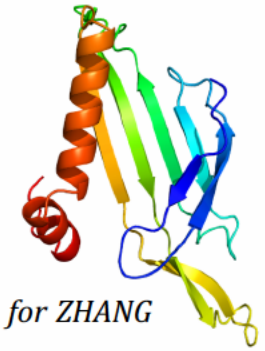
ZHANG Model1

TM-score=0.86
RMSD=2.26Å

Zhang-Server Model1



TM-score=0.62
RMSD=10.2Å



TM-score=0.81
RMSD=3.03Å
Initial conformation for ZHANG

ZHANG vs. Zhang-Server

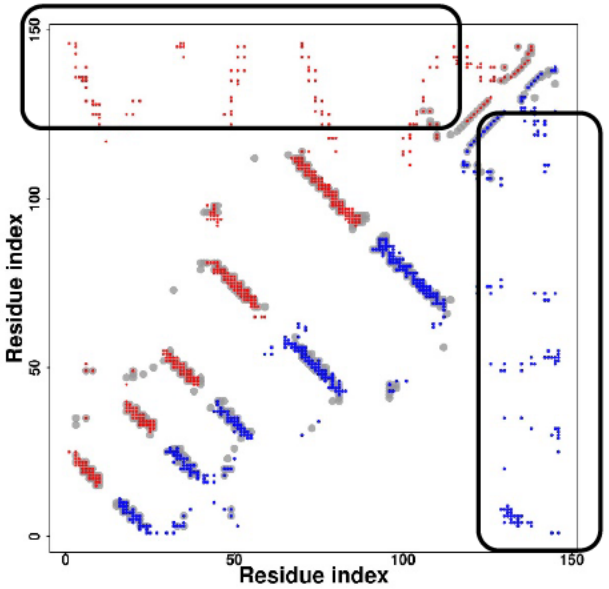
Experimental

CASP:
Images redacted

Zhang-Server Model1

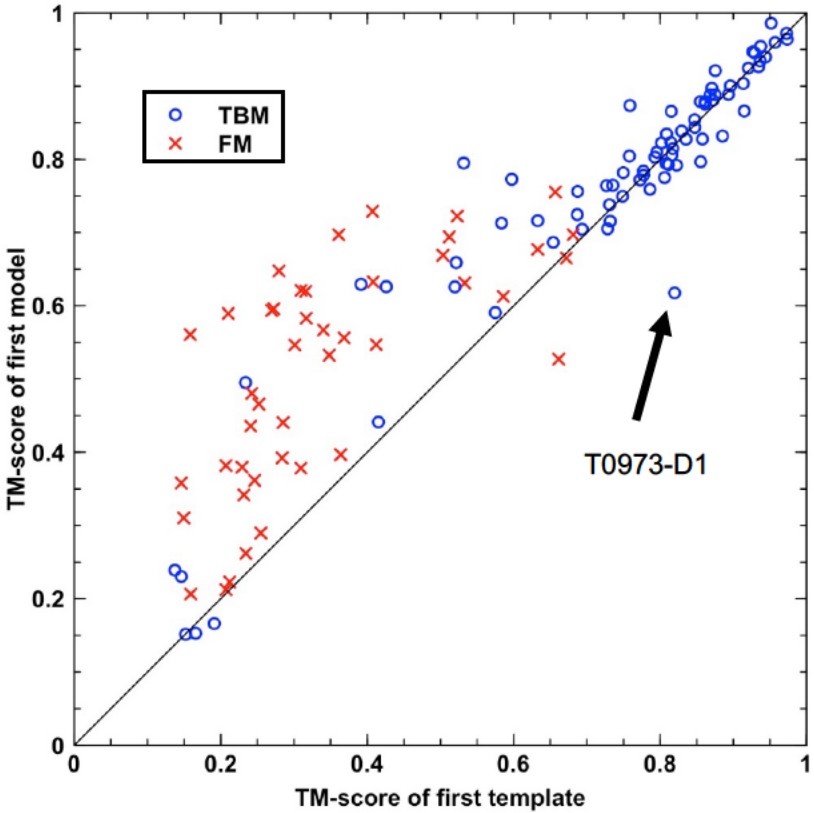
TM-score=0.62
RMSD=10.2Å

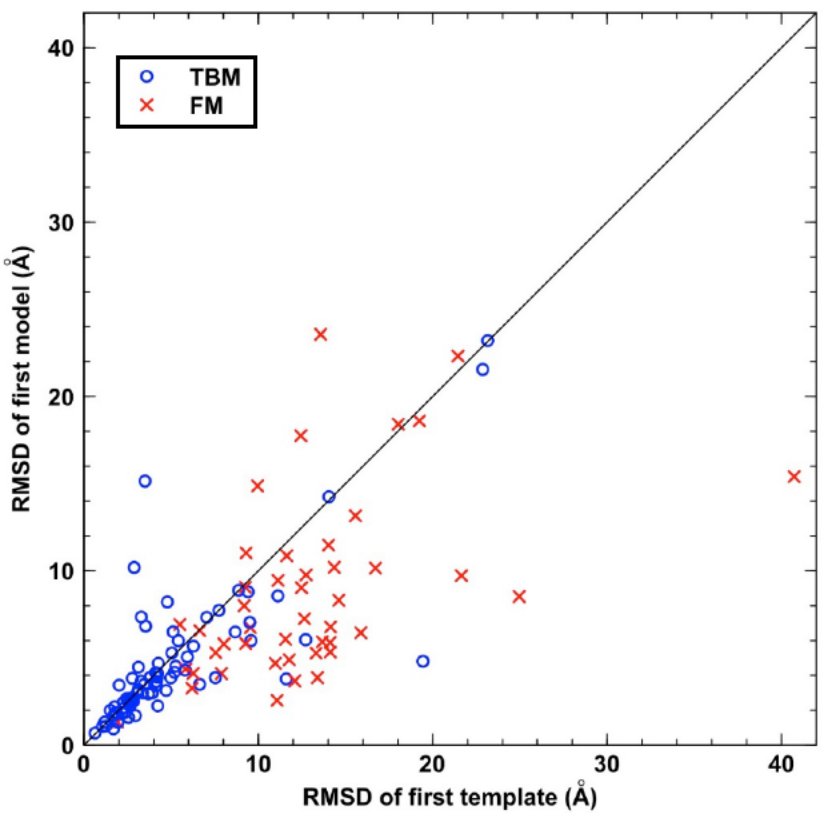
Experimental (rainbow)
Template (black)
TM-score=0.82
RMSD=2.88Å



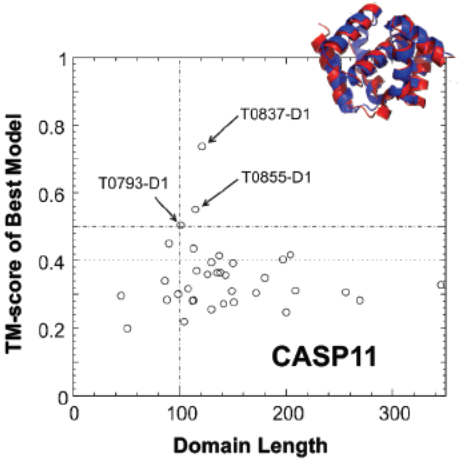
Good template, bad long range contacts!

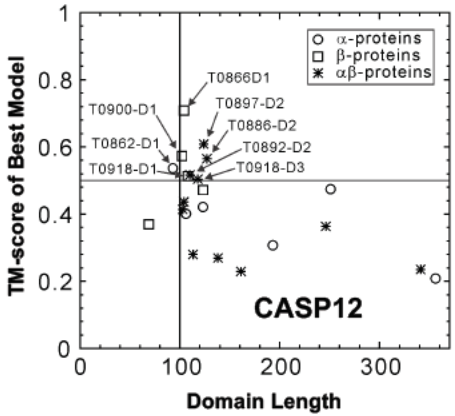
First Zhang-Server Model vs. First LOMETS Template)



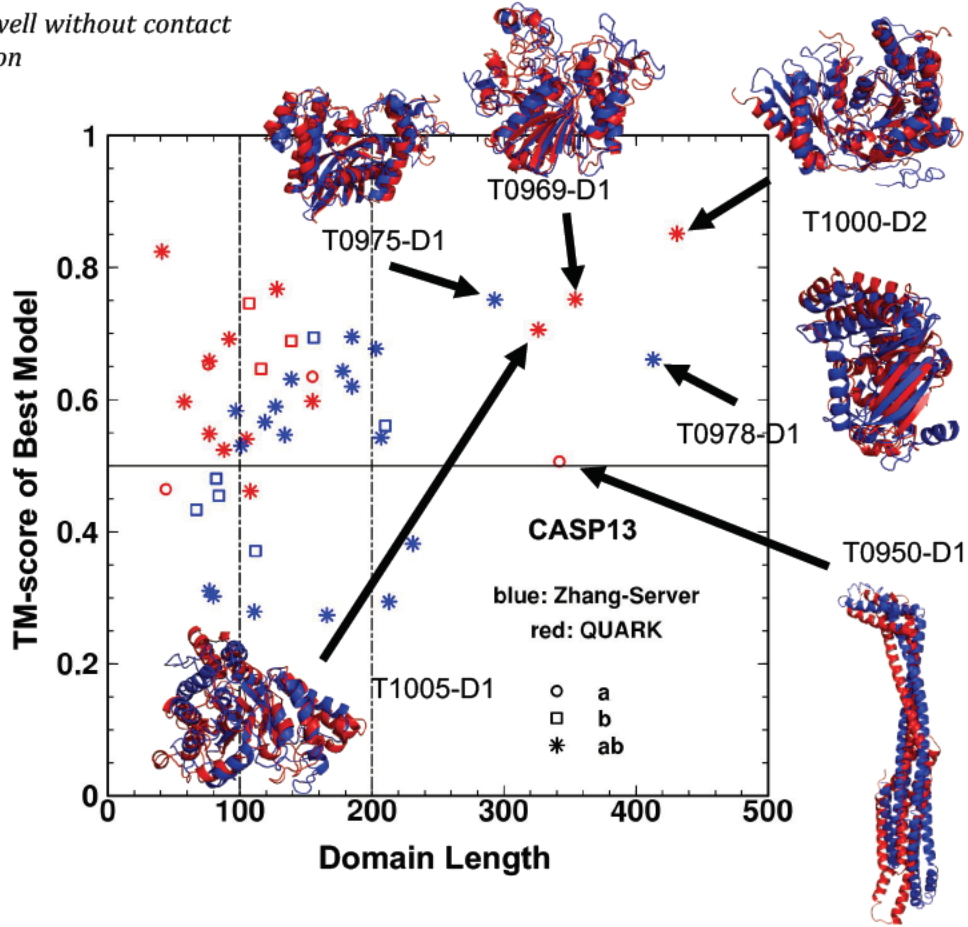


Summary of 45 FM targets (32 FM targets, 13 FM/TBM targets)





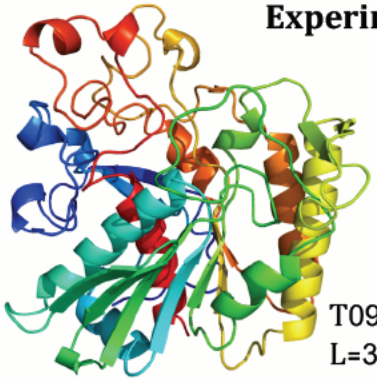
Folded well without contact prediction



Experimental Structure (Red)

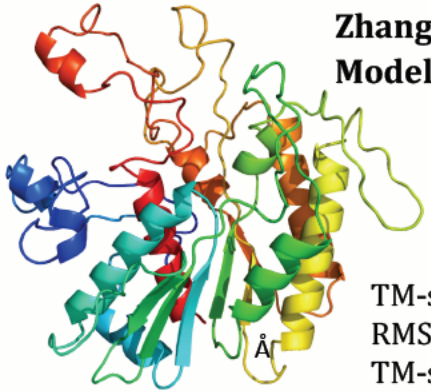
Model (Blue)

Folding Large Targets



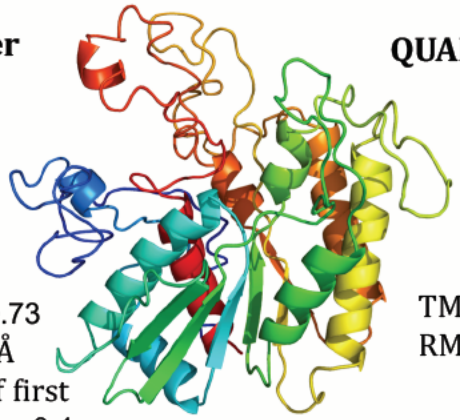
Experimental

T0969-D1
L=354AA
Long range top L: 70.06%



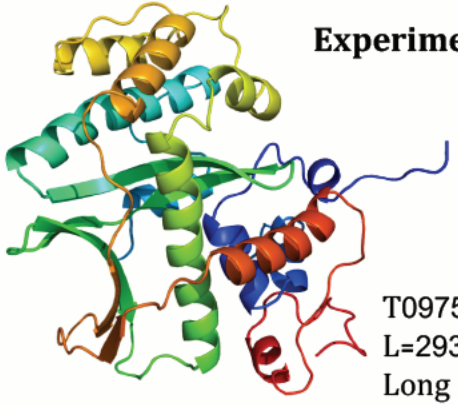
Zhang-Server Model1

TM-score=0.73
RMSD=6.08Å
TM-score of first template is ~ 0.4



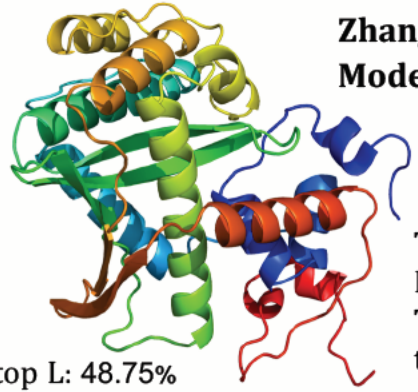
QUARK Model1

TM-score=0.75
RMSD=5.41Å



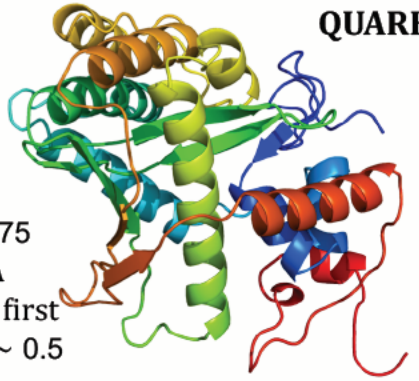
Experimental

T0975-D1
L=293AA
Long range top L: 48.75%



Zhang-Server Model5

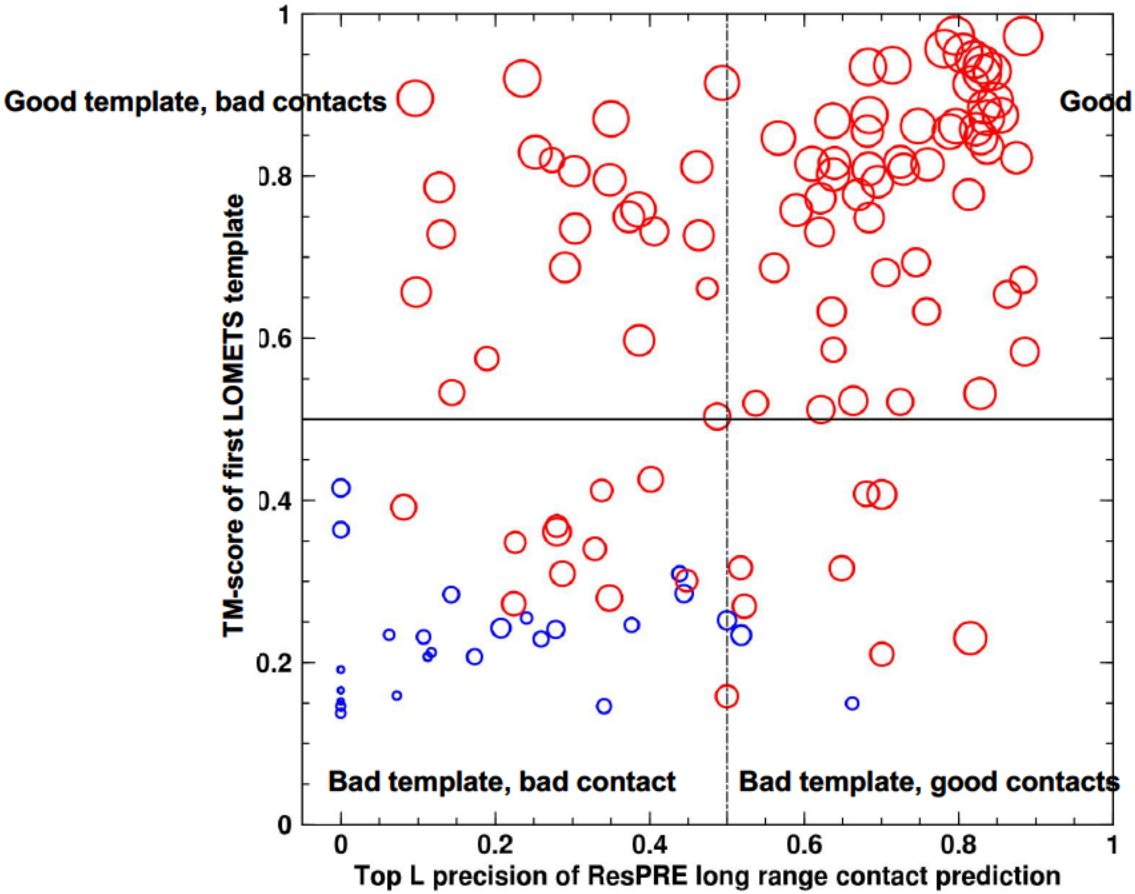
TM-score=0.75
RMSD=4.95Å
TM-score of first template is ~ 0.5



QUARK Model5

TM-score=0.74
RMSD=5.02Å

Impact of Contact and Template Quality on Structure Prediction

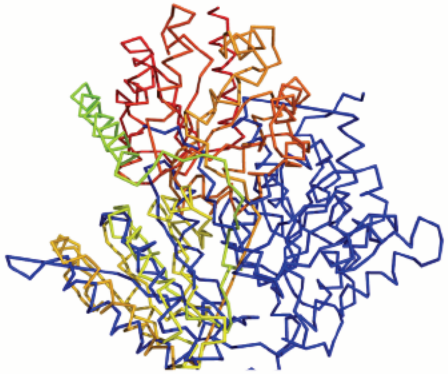


Red: foldable
(TM-score of model >= 0.5)

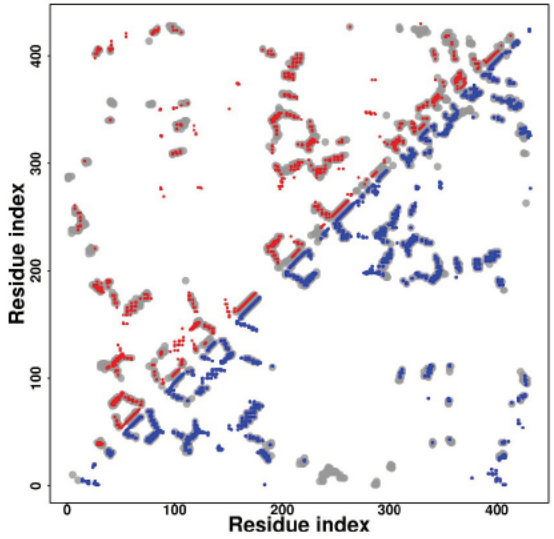
Blue: not foldable
(TM-score of model < 0.5)

Larger marker = higher TM-score

T1000-D2 (FM Target)
(Bad Template, Good Contacts)



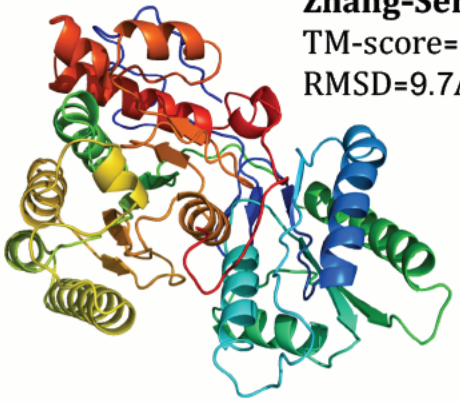
Best template: 1jqkA
TM-score=0.21
RMSD=21.65Å
L=431AA



Grey: Native contacts
Red: ResPRE contacts
Blue: Zhang-Server model1 contacts

ResPRE Long range top L: 81.52%

Experimental

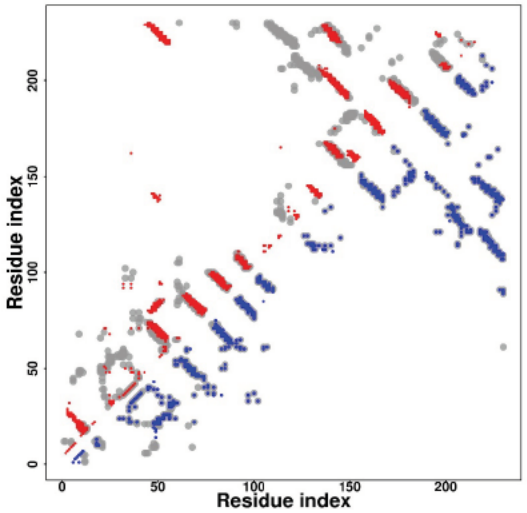


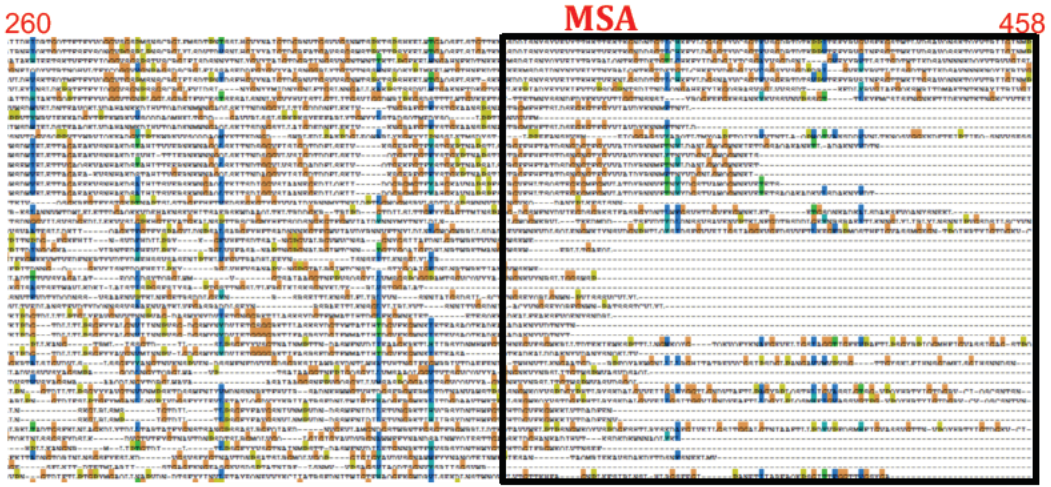
Zhang-Server Model1
TM-score=0.814
RMSD=9.7Å

QUARK Model1
TM-score=0.851
RMSD=4.29Å

T1004-D3 (Good Template, Bad Contacts)

Sparse MSA for contact prediction but good structure template exists!



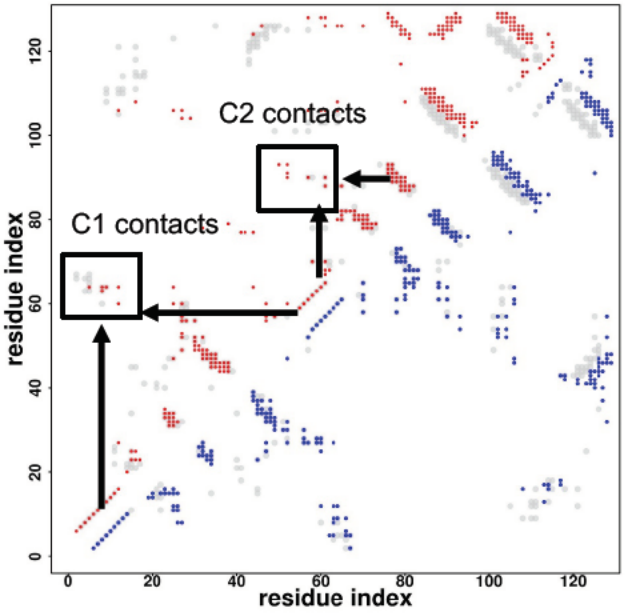


CASP:
Image redacted

T1004-D3: 229-458
Template: 5m9fA, TM-score=0.92,RMSD=2.02Å
Long range, top L=23.48%
Zhang-Server Model1 TM-score=0.925

T1017s2 (Bad Template, Bad Contacts, Good Fold)

CASP:
Image redacted



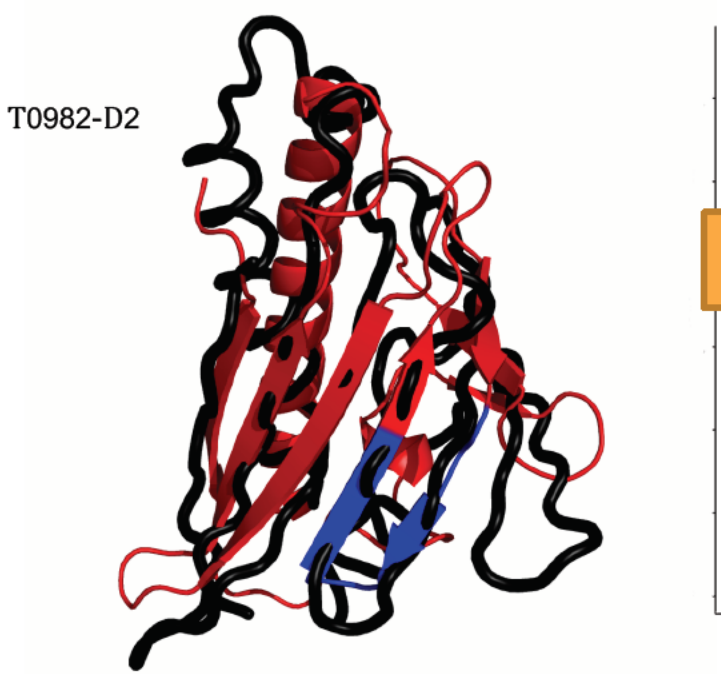
CASP:
Image redacted

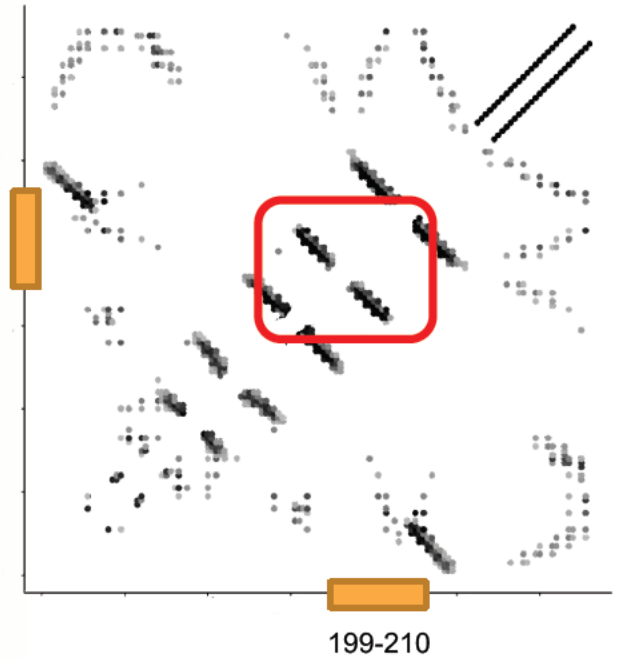
Template: 5m9fA (yellow)
TM-score=0.36
RMSD=6.3Å
ResPRE Long range, top L=28%

Balance template and contacts!

Zhang-Server Model1
TM-score=0.70
RMSD=4.1Å
ResPRE Long range, top L=28%

Secondary Structure Prediction Problem






```
Conf: 885279857997404452447988989862224757960698778874888877558971
Pred: CCCEEEEECCCCCCCEEEEHHHHHHHHHHHHHCCEEECCCCEEEEEEEECCCCCCCCEEEEE
  AA: GPGWFFVVDEEGPRRFTLRDWQLDRERALTFAVEIPGARTVTACQVRTEPGERGRTLSVS
      190       200       210       220       230       240
```

Domain Partition Problem

Experimental

T1011-D2

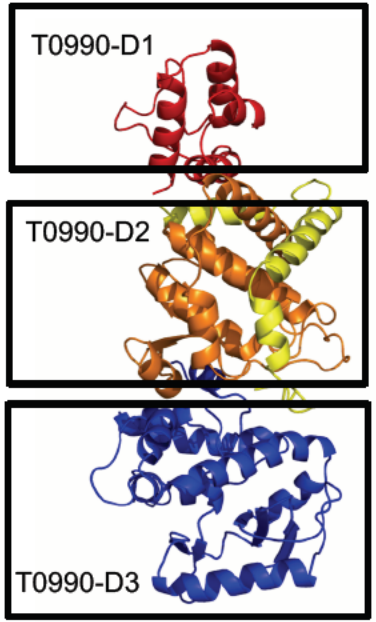
T1011-D1

Zhang-Server model1

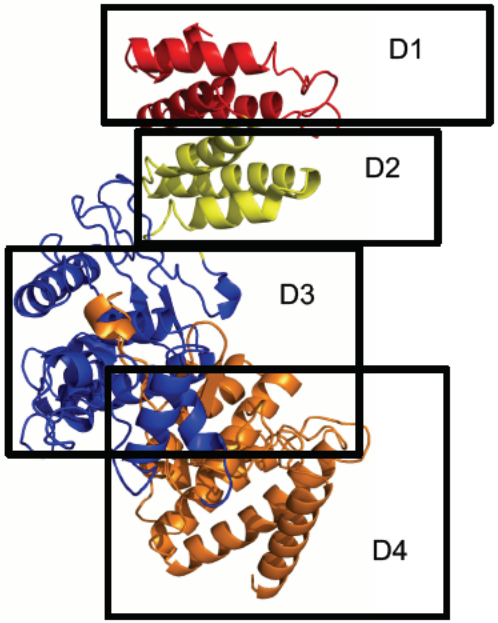
D1: TM-score=0.81
D2: TM-score=0.87

CASP domain: D1:55-268,433-520 D2: 271-430
Zhang-Server domain: D1:8-267,434-519 D2:268-433

Experimental Domain Partition

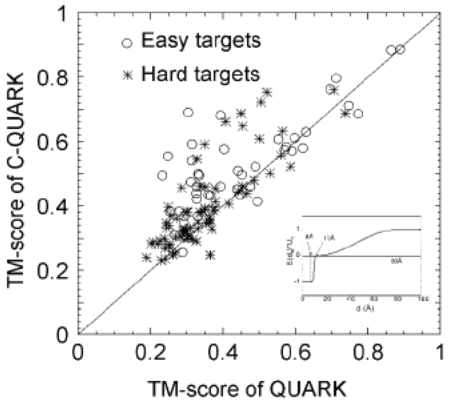


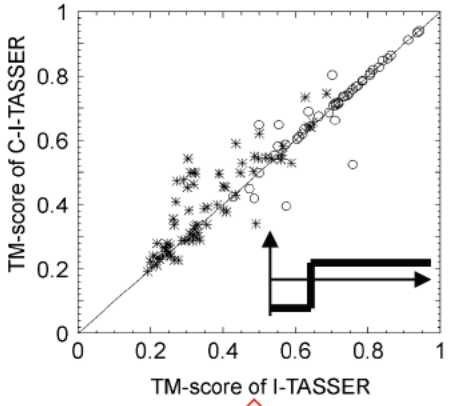
T0990-D1 TM-score=0.57
T0990-D2 TM-score=0.38
T0990-D3 TM-score=0.21



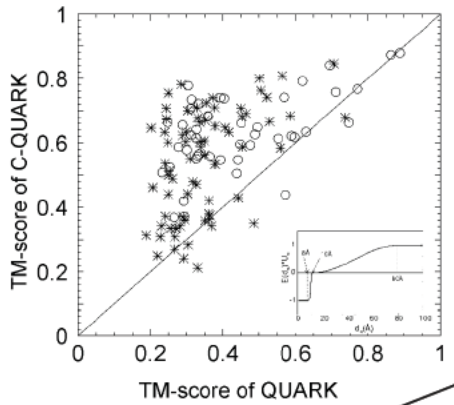
CASP domain: D1:1-76 D2:77-134,348-520 D3: 135-347
Zhang-Server domain: D1:1-128 D2:129-280 D3:281-406 D4: 407-552

Impact of Contact Maps on 3D Structure Prediction



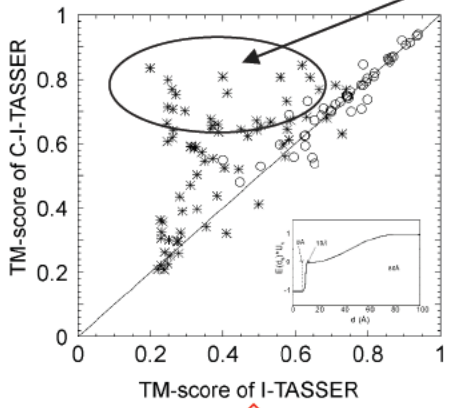


CASP12 (SVMSEQ+NeBcon)



QUARK vs C-QUARK

Many non-homology proteins can now fold as well as homology proteins



I-TASSER vs C-I-TASSER

CASP13 (ResPRE+NeBcon)

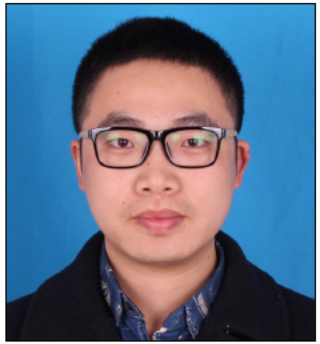
Summary

- What worked
  - Contact prediction by Deep Neural Networks
  - Integration of contact potential, LOMETS restraints, and inherent potential
- What needs improvement
  - Secondary structure prediction
  - Discontinuous domain partitioning

Acknowledgements

Wei Zheng

Chengxin Zhang



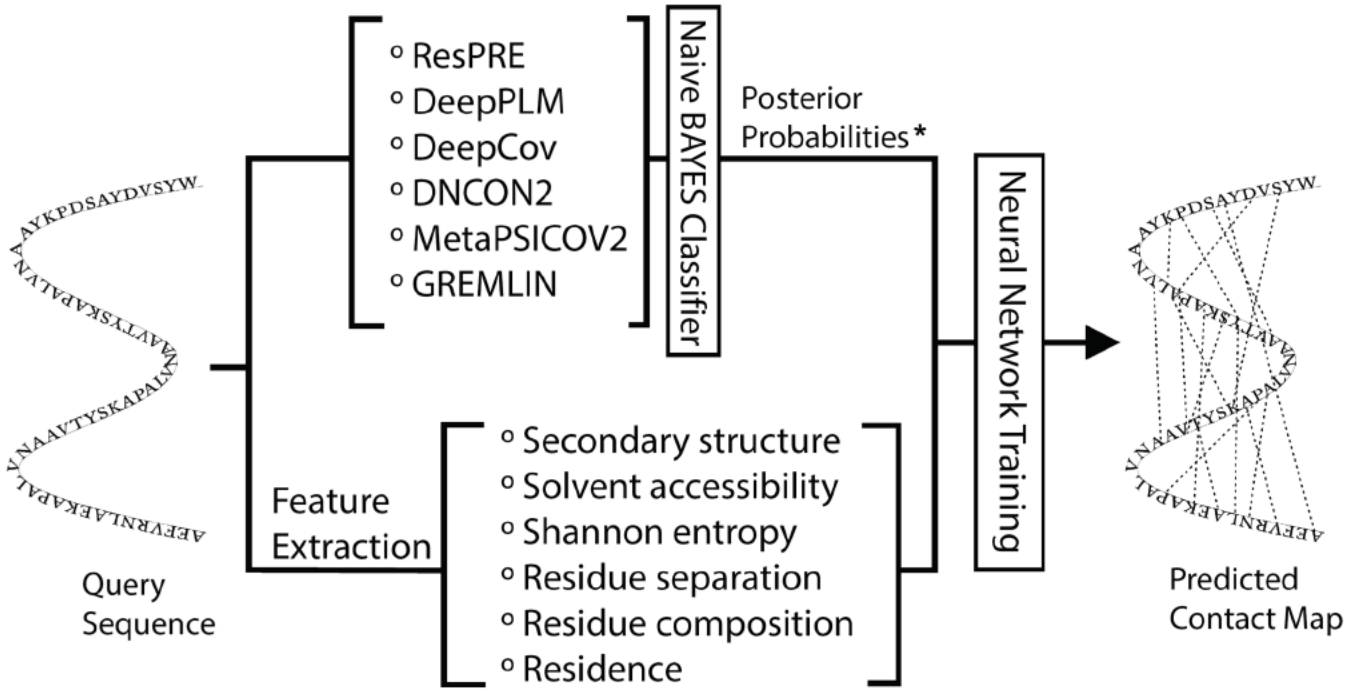
Yang Li

S M Mortuza

Yang Zhang

- Experimentalists that shared their structures
- CASP Organizers

Contact Prediction using NeBcon



$$*P\left(C\middle|X_{ij}^{1},X_{ij}^{2},\cdots,X_{ij}^{N}\right)=\frac{P(C)\prod_{m=1}^{N}P\left(X_{ij}^{m}\middle|C\right)}{P(0)\prod_{m=1}^{N}P\left(X_{ij}^{m}\middle|0\right)+P(1)\prod_{m=1}^{N}P\left(X_{ij}^{m}\middle|1\right)}$$