Tertiary structure (TS) prediction and refinement from Baker groups

Ivan Anishchenko, Minkyung Baek, Hahnbeom Park, Justas Dauparas, Naozumi Hiranuma, Sanaa Mansoor, Ian Humphrey, and David Baker
Groups from Baker Lab

- **BAKER**
  - server pipeline +
  - metagenomes +
  - refinement

- **BAKER-experimental**
  - server pipeline +
  - metagenomes +
  - assemblies

- **BAKER-ROSETTASERVER**
  - new DL-based server

- **BAKER-ROBETTA**
  - TBM +
  - “classical” coevolution +
  - refinement

* has not changed since CASP13 (&12)
BAKER-ROSETTASERVER

Target sequence

hhblits, Uniclust30

hhsearch, PDB100

trRosetta

trRosetta w/ templates

Fold by minimization & relax

PyRosetta

Rescore & recombine

DeepAccNet

trRefine

Fold by minimization, relax & rescore

PyRosetta, DeepAccNet

Models 1-3

Model 4

Model 5

2D ResNets

takes several hours for an average size protein

BAKER-ROSETTASERVER
human interventions

more sequences

domain splitting/merging,
modeling in the context of an oligomer
(BAKER-experimental, talk on Wed)

large-scale refinement

Target sequence

hhblits, Uniclust30

hhsearch, PDB100

trRosetta

trRosetta w/ templates

Fold by minimization & relax

PyRosetta

Rescore & recombine

DeepAccNet

trRefine

Fold by minimization, relax & rescore

PyRosetta, DeepAccNet

Models
trRosetta
_transform-restrained Rosetta

input MSA
sequence, PSSM, entropy
couplings, seq. separation

2D ResNet

residue-residue geometries (6 DoF)

binned predictions @ d<20Å

J Yang, I Anishchenko, H Park, Z Peng, S Ovchinnikov, D Baker,
*Improved protein structure prediction using predicted interresidue orientations*, PNAS, 117: 1496-1503 (2020)
### trRosetta
(transform-restrained Rosetta)

- **input MSA**
- **sequence, PSSM, entropy**
- **couplings, seq. separation**

**2D ResNet**

**residue-residue geometries (6 DoF)**

**Fold by minimization & relax**

- **PyRosetta**
- **45 models**

**3D model**

- **binned predictions @**
  - **d**: 0 - 20 Å
  - **ω**: -180° - 180°
  - **θ**: 0 - 180°
  - **φ**: 0 - 180°
**trRosetta** with templates

**Input MSA**

- sequence, PSSM, entropy
- couplings, seq. separation

2D ResNet

**Features from templates**

- conv2d
- conv2d
- conv2d

**“pixel-wise” attention**

- template 1
- template 2
- template 25

**HHsearch hit example**

- Proba=5.04
- E-values=3.2e-03
- Score=26.24
- Aligned_cols=24
- Identities=33
- similarity=0.662

<table>
<thead>
<tr>
<th>Q ss_pred</th>
<th>cecCcE----ECHHHHHHHHcCCCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q T1052</td>
<td>B09 GLYGAK-----VNAFIA0FKSKGWGG</td>
</tr>
<tr>
<td>Q Consensus</td>
<td>B09 ~--------------w--f--f--f--k--w</td>
</tr>
<tr>
<td>T Consensus</td>
<td>110 vpyssvk-khr---+e-v-emky+y+feskwp-</td>
</tr>
<tr>
<td>T lo22 A</td>
<td>110 VPyssvkKKKRRRLVEFPMKFYFFESKGNWP</td>
</tr>
<tr>
<td>T ss_dssp</td>
<td>EEEGGSTTCCHHHHHHHHHHHHHHHTCCCG</td>
</tr>
<tr>
<td>T ss_pred</td>
<td>EHHHccccHHHHHHHHHHHHHHHHhccc</td>
</tr>
<tr>
<td>Confidence</td>
<td>00111111 2356666666677774</td>
</tr>
</tbody>
</table>

**HHsearch stats**

- similarity
- confidence
- d, ω, θ, φ, hhsearch stats
network with templates typically gives better predictions than the MSA-only net

in low $N_{eff}$ regime unrelated templates may misguide predictions

merge predictions from both networks

$trRefine$
Model rescoring and recombination

trRosetta predictions

10 top scored trRosetta models & predicted errors by DeepAccNet

Refined 6DoF predictions

trRefine (2D ResNet)

Folding (pyRosetta)

Refined model
T1055-D1, model1 (ΔGDT-TS = +6.4)
Recombining models with *trRefine* generally improves their quality.
Submitted models are generally of better quality than the best selected \textit{hhsearch} template ...

... but some good templates were missed

![GDT_TS Diagram](image)

![GDT_TS Diagram](image)
Prediction workflow by the example of T1079

**MSA for T1079**

- **trRosetta**
  - GDT_TS = 63.75
- **GDT_TS = 68.29**
- **next best server GDT_TS = 63.69**

**top 25 hhsearch hits**

- **trRosetta w/ templates**
  - GDT_TS = 68.29
- **trRefine**
  - GDT_TS = 72.12
Recombination of templates

features from templates

"pixel-wise" attention

cov2d 1

cov2d 2

cov2d ...

template 1

template 2

template ...

template 25
Recombination of templates

T1052

features from templates

“pixel-wise” attention

conv2d

template 1

template 2

... template 25

green bars

residues 81-251

contribution

0.25
0.20
0.15
0.10
0.05
0.00

0 10 20

template

top 25 hhsearch hits

GDT_TS = 75.60

best tmplt = 78.15
Recombination of templates

T1052

GDT_TS = 75.60
best tmplt = 78.15
next best ΔGDT_TS = -4.11

features from templates

“pixel-wise” attention

conv2d

residues 81-251

residues 571-771

residues 621-661

top 25 hhsearch hits
templates were missed

no good templates in top 25

plenty of templates down the list

used for network predictions

could have been modeled better

<table>
<thead>
<tr>
<th></th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
</tr>
</thead>
<tbody>
<tr>
<td>best tmplt, top25</td>
<td>46.04</td>
<td>21.96</td>
<td>21.70</td>
<td>8.71</td>
</tr>
<tr>
<td>best tmplt, top500</td>
<td>46.04</td>
<td>61.92</td>
<td>72.41</td>
<td>78.57</td>
</tr>
<tr>
<td>model1</td>
<td>76.08</td>
<td>77.34</td>
<td>65.80</td>
<td>43.97</td>
</tr>
</tbody>
</table>

T1091: top hhsearch hits

hhsearch probability

0 100%
server did poorly on hard targets ...

\[ Z_{\text{TS}} + Z_{\text{QCS}} \]

... but much better on easier ones

\[ Z_{\text{HA}} + \frac{Z_{\text{SG}} + Z_{\text{IDDT}} + Z_{\text{CAD}}}{3} + Z_{\text{ASE}} \]
Server Summary

- Joint use of templates and MSA worked well
- Templates could have been selected and used better
- No good reason for not using more sequences (metagenomes)
- *trRefine* does improve model quality but not dramatically
Human Tertiary structure prediction & Refinement (BAKER)
Human vs Server:
Contribution from Additional Sequence search for trRosetta

Human modeling
- **26 domains**: Submitted as server models
- **67 domains**: Remodeled with alternative MSAs
  (open circles: starting points for refinement)
Example 1: Sequence search & Modeling as a whole Protein

Server: per-target MSA, Nseq=1~3 (UniRef30)
Human: MSA & modeling on entire protein, Nseq>2000 (+MetaGenome, IMG/VR DB)
Example 2: Chimeric distance map for multi-domain targets

T1085-D1 (gray)

T1085-D2 (magenta->cyan)
GDT-TS: 60 -> 83

Poor sequence coverage

Separate sequence search on D2

Distance map on the entire length
Refinement guided by DeepAccNet

Key idea of refinement in CASP14: To use EMA to guide Refinement search

Signed distance error predictions from DeepAccNet

N Hiranuma et al, Improved protein structure refinement guided by deep learning based accuracy estimation, bioRxiv.
Refinement protocol

Iterative Rosetta+DeepAccNet refinement
Refinement results (colored by size)

- TS (52 domains)
- TR (44 targets)

- Refinement steps:
  - GDT-HA + 4.4
  - GDT-HA + 2.0

- Open circles: part of the large viral protein

- Protein size range: 80 to 220
What went wrong -- R1035  $\Delta$(GDT-HA): -7

Xtal-structure

Starting/Refined as domain

T1035: Starting/Refined together with T1033
What went wrong -- R1038-D2

Δ(GDT-HA): -6

Xtal-structure

Refined as domain

Inter-domain hydrophobic contacts
What went wrong -- R1078

$\Delta$(GDT-HA): -15

Xtal-structure (homo-dimer)

Chain B

Starting/Refined as monomer

Broken Disulfide bond

Dimer interface
**Bright side:** EMA-guided refinement enables improving relatively **larger proteins**
Still challenging -- R1068, R1056, etc.

Could not improve significantly when topologies were complicated & sizes were big
Take home messages

- EMA-guided refinement tested in CASP14; general challenge remains

- Refinement in a monomeric context fails; Information is often more critical than principles in real practice modeling scenarios
Acknowledgements

BAKER group CASP14 team
Ivan Anishchenko
Minkyung Baek
Hahnbeom Park
David Kim
Justas Dauparas
Naozumi Hiranuma
Sanaa Mansoor
Ian Humphrey
Luki Goldschmidt
David Baker

trRosetta developers
Jianyi Yang
Sergey Ovchinnikov

CASP organizers & participants