Protein Assembly by Server and Human

Taeyong Park, Hyeonuk Woo, Jinsol Yang, Sohee Kwon, Jonghun Won and Chaok Seok
Contents

1. Seok assembly pipeline
2. Server vs. Human
3. What went right (four targets)
4. Deep learning (DL) and protein assembly
Seok assembly pipeline
Server pipeline on $A_n$ target

- Sequence
  - CASP14 TS protocol
  - A monomer model
    - Sequence-based template search & TBM
    - Structure-based template search & TBM
      - Ab initio docking by GalaxyTongDock_C
    - Oligomer models
      - ULR modeling by GalaxyLoop (Model1)
      - Refinement by GalaxyRefineComplex (Model1)

5 models
Server pipeline on $A_n$ target

CASP14 TS protocol

Sequence

Sequence-based template search & TBM

Structure-based template search & TBM

Ab initio docking by GalaxyTongDock_C

Oligomer models

ULR modeling by GalaxyLoop (Model1)

ML-based selection

A monomer model

5 models

Sequence

TBM

DBM
Server pipeline on $A_1B_1$ target

- **Sequences of A & B**
  - Template search for A & B using HHsearch
    - Common templates of A and B
      - Monomer templates
      - Homo-oligomer templates
      - Hetero-dimer templates
  - CASP14 TS protocol
    - Hetero-dimer template search
    - Monomer structures of A & B
  - Modeling AB by superposition on the found templates
  - Ab initio docking by GalaxyTongDock_A

- Oligomer models
  - ULR modeling by GalaxyLoop (Model1)
  - Refinement by GalaxyRefineComplex (Model1)

- 5 models
Server pipeline on $A_1B_1$ target

1. **Sequences of A & B**

2. **Template search for A & B using HHsearch**
   - Common templates of A and B
   - **Monomer templates**
   - **Homo-oligomer templates**
   - **Hetero-dimer templates**

3. **CASP14 TS protocol**
   - **Hetero-dimer template search**
   - Monomer structures of A & B

4. **Modeling AB by superposition on the found templates**

5. **Ab initio docking by GalaxyTongDock_A**

6. **Oligomer models**

7. **ULR modeling by GalaxyLoop (Model1)**

8. **Refinement by GalaxyRefineComplex (Model1)**

9. **5 models**
Server prediction on the targets of more complicated stoichiometry ($A_l B_m C_n ...$)

- Individual interfaces were predicted by GalaxyHeteromer and GalaxyHomomer2.

- In the case of H1072 ($A_2B_2$), $A_2$ and $B_2$ interfaces were predicted by GalaxyHomomer2. AB interface was predicted by GalaxyHeteromer.
1. Available information from the literature and human insight were employed for model generation and selection.

2. Monomer models from other servers were also utilized to predict oligomer structures.
Server vs. Human
Server vs. Human

![Graph showing the comparison of Server vs. Human top 5 best Fnat values. The graph includes data points for Homo-oligomer (blue) and Hetero-oligomer (orange).]

- Homo-oligomer data points: T1070, T1083, T1078, T1099v1
- Hetero-oligomer data points: T1083

Key:
- **Homo-oligomer**
- **Hetero-oligomer**
What went right
(four targets)
1. T1070
(\textit{Ab initio docking})
T1070 \((A_3)\)
Manually separated into 4 domains based on server models for monomer.

- **domain 1**: 1-76
- **domain 2**: 80-165
- **domain 3**: 190-249
- **domain 4**: 264-335

Crystal Model
The crystal structure also can be divided into similar 4 domains:

- **Domain 1**: 1-76 (1-76)
- **Domain 2**: 80-180 (80-165)
- **Domain 3**: 181-256 (190-249)
- **Domain 4**: 265-335 (264-335)
The crystal structure also can be divided into similar 4 domains

- **Domain 1:** 1-76 (1-76)
- **Domain 2:** 80-180 (80-165)
- **Domain 3:** 181-256 (190-249)
- **Domain 4:** 265-335 (264-335)

T1070 (A₃)
T1070 (A$_3$) domain 2 (80-180)

Model 5: little bit closer each other because of lack of 166-180 due to mis-spliting of the domain
T1070 (A₃) domain 2 (80-180)

Model 5: little bit closer each other because of lack of 166-180 due to mis-spliting of the domain

Because we defined domain 2 as 80-165, 166-180 is missing

Model 5

$F_{nat}$: 0.18
IRMSD: 5.26Å
LRMSD: 5.31Å

“Acceptable quality”
T1070 (A₃) domain 2 (80-180)

Model 5: little bit closer each other because of lack of 166-180 due to mis-splitting of the domain

Because we defined domain 2 as 80-165, 166-180 is missing

Model 5

\[ F_{\text{nat}}: 0.18 \]
\[ \text{IRMSD}: 5.26\text{Å} \]
\[ \text{LRMSD}: 5.31\text{Å} \]

“Acceptable quality”

GalaxyTongDock top 1 prediction by using Seok-server_TS1 monomer structure

Scoring based on human intuition didn’t work
T1070 (A₃) domain 4 (265-335)
Domain-split was successful 265-335 vs. 266-335.

Model 1
$F_{nat}$: 0.42
IRMSD: 3.22Å
LRMSD: 4.06Å

"Medium quality"

GalaxyTongDock top 16 prediction by using Zhang-CEthreader_TS1
(The best docking pose generated)

Scoring based on human intuition worked well
2. T1078

(\textit{Ab initio docking})
T1078 (A₂)
Model 3 is one of the four acceptable quality models.
T1078 (A$_2$)

Model 3 is one of the four acceptable quality models.

$F_{\text{nat}}$: 0.18
IRMSD: 4.28Å
LRMSD: 6.95Å
T1078 (A₂)
Model 7 has right orientation, but is of lower quality.

The two termini affected the model quality.

F_{nat}: 0.15
IRMSD: 5.50Å
LRMSD: 7.80Å

“Acceptable quality”
T1078 (A₂)

Model 7 has right orientation, but is of lower quality.

The two termini affected the model quality.
T1078 (A2)

Unsuccessful N-terminal modeling (1-14)

GalaxyTongDock was used with only 15-128 region.

Crystal s1
Model s1
Crystal s2
Model s2

$F_{nat}: 0.15$

$IRMSD: 5.50\text{Å}$

$LRMSD: 7.80\text{Å}$
T1078 (A₂)

Unsuccessful N-terminal modeling (1-14)

GalaxyTongDock was used with only 15-128 region.

Evaluation without N-terminal (1-14)

- \( F_{\text{nat}} \): 0.15
- \( \text{IRMSD} \): 5.50 Å
- \( \text{LRMSD} \): 7.80 Å

Crystal s1
Model s1
Crystal s2
Model s2
T1078 (A$_2$)
Hexahistidine tag (LPLEHHHHHH, 129-138)

Part of the exp tag forms strong interaction.

Evaluation without N-terminal (1-14)
$F_{nat}$: 0.25
$IRMSD$: 1.72Å
$LRMSD$: 3.65Å
T1078 (A2)
Hexahistidine tag (LPLEHHHHHH, 129-138)

We didn’t model the expression tag.

Evaluation without N-terminal (1-14)
$F_{nat}$: 0.25
IRMSD: 1.72 Å
LRMSD: 3.65 Å

Also without exp tag (129-138)
$F_{nat}$: 0.43
IRMSD: 1.72 Å
LRMSD: 3.69 Å

“Medium quality”

Monomer quality (FEIG)
RMSD: 1.63 Å
3. T1083
(Template-based docking and refinement)
Model 1

$F_{nat}$: 0.6078
IRMSD: 2.51 Å
LRMSD: 4.83 Å

Monomer quality
(Seok_assembly)
RMSD: 2.82 Å

The only "Medium quality" model by server

Crystal
Model s1
Model s2
T1083 (A$_2$)

Monomer model

Monomer quality (Seok _assembly)
RMSD: 2.82Å

Structure-based template

Superposed model s1
Superposed model s2

Relaxed structure

Relaxed model s1
Relaxed model s2
T1083 (A₂)

Before refinement:

\( F_{nat}: 0.3922 \)
\( IRMSD: 4.09\text{Å} \)
\( LRMSD: 10.147\text{Å} \)

“Incorrect quality”

Crystal Model s1
Model s2

After refinement:

\( F_{nat}: 0.6078 \)
\( IRMSD: 2.51\text{Å} \)
\( LRMSD: 4.83\text{Å} \)

“Medium quality”

GalaxyRefineComplex
4. T1099
(Template-based docking and data-assisted refinement)
T1099v1 (A₂)

Model 2

\[ F_{\text{nat}}: 0.3398 \]
\[ IRMSD: 3.90\text{Å} \]
\[ LRMSD: 8.04\text{Å} \]

The only “Acceptable quality”

Monomer modeling starts from (FALCON-TBM TS1) RMSD: 9.5Å
T1099v1 (A₂)
Relative orientation & insertion region for DHBc.

Template (PDB ID:3J2V) is available. seqID*coverage = 12.1.

Template (HBV) vs Crystal (DHBc)
- Long insertion region at DHBC
- Different relative orientation of main helix.

From literature search, information about the role and binding mode of insertion region for assembly was obtained.
T1099v1 (A₂)

Insertion region for DHBc.

Crystal structure

Our model

Fₙₐᵗ: 0.3398
IRMSD: 3.90 Å
LRMSD: 8.04 Å

“Acceptable quality”

Monomer modeling starts from (FALCON-TBM TS1) RMSD: 9.5 Å

Crystal s1
Crystal s2
Model s1
Model s2
A preliminary study of applying deep learning to ab initio docking
Re-scoring docking poses by deep learning

• **GalaxyTongDock**, a grid-based rigid-body docking, has limited scoring power.

  ![Diagram of GalaxyTongDock](image)

• Deep learning-based re-scoring method has been developed.
Simple interface scoring by deep learning

- Residue of interest
- Ten nearest residues from partner chain
  *(within 11 Å by Cb-Cb distance)*

ML

Per-residue interface score

Residue-sum

Interface score for each pose
Deep learning-based re-scoring improves performance of *ab initio* docking

Non-redundant 1,568 hetero-dimers
> 1,156 (training set) and 412 (test set)

### Performance on 412 hetero-dimers in test set

<table>
<thead>
<tr>
<th></th>
<th>TongDock</th>
<th>DL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Top 1</strong></td>
<td>0.49</td>
<td>0.73</td>
</tr>
<tr>
<td><strong>Top 10</strong></td>
<td>1.70</td>
<td>1.70</td>
</tr>
<tr>
<td><strong>Top 100</strong></td>
<td>2.91</td>
<td>2.91</td>
</tr>
<tr>
<td><strong>Top 1,000</strong></td>
<td>3.40</td>
<td>4.37</td>
</tr>
</tbody>
</table>

Where

- *** indicates p-value < 0.001
- ** indicates p-value < 0.01
- * indicates p-value < 0.05
Take-home message

1. *Improved monomer quality improved performance of ab initio docking.*

2. *Interface refinement can improve oligomer model quality.*

3. *Advent of deep learning is changing everything.*
Thank you for listening