Assessment of Template Based Modeling (TBM)

CASP7 Meeting
Asilomar, November 28, 2006

Torsten Schwede
Swiss Institute of Bioinformatics
Biozentrum, University of Basel
Acknowledgements

Jürgen Kopp
Lorenza Bordoli
Florian Kiefer
James Battey

- Andriy Kryshtafovych
- Randy Read
- Neil Clarke
- Michael Tress

- Anna Tramontano
- Roland Dunbrack
- Alfonso Valencia

- John Moult
- Krzystof Fidelis
- Burkhard Rost
- Tim Hubbard

- Ernest Feytmans
Overview

1. TBM targets and models in numbers
2. Numerical assessment
   - GDT / AL0 Z-Scores
   - Assessment of local atomic interactions
3. Results of the assessment
   - Statistical significance and ranking
   - Specific questions to predictors
4. Progress & Improvement
   - Improvement over template
   - Comparison with CASP6
5. Assessment of model confidence ("B-Factors")
108 domains for TBM assessment

28 HA-TBM target domains
Target domains for which suitable template structures (LGA score > 80) were available, and the best model has at least GDT_TS of 80.

76 TBM target domains
Target domains for which at least one structurally similar template was available (LGA vs. PDB), and a template has been used in at least one prediction.

4 TBM/FM overlap domains
Target domains for which a significant fraction of the secondary structure elements could not be modeled with correct topology based on a single template (by visual inspection).
Assessment of TBM (Template Based Modeling)

- 187 predictor groups in the assessment
  - 68 servers
  - 119 expert groups

- 15’711 “No. 1” target domain predictions > 20 residues
  - 14’636 TS domain models
  - 1’075 AL domain models
Most groups predicted > 100 domains.

Only few groups with less than 20 predictions.
CASP7 – Number of predictions submitted per target domain

Target domains sorted by number of submitted predictions

T0383 (TBM): 145 predictions
T0316_D3 (TBM): 101
T0295_D1/2 (HA-TBM): 102
Assessment of models

- Groups with less than 20 predicted domains were assessed, but not included in the final ranking statistics.

- Models with severely unrealistic geometry:
  - 451 models flagged as physically impossible as defined:
    - > 2% of C\(\alpha\) atoms in clashes: \(0 < d(C\alpha-C\alpha) < 1.9 \, \text{Å}\) or
    - > 10% of C\(\alpha\) atoms in bumps: \(1.9 < d(C\alpha-C\alpha) < 3.5 \, \text{Å}\) or
    - unrealistic models by visual inspection (e.g. atom compressions, severe fragmentation).
  - Unfeasible models were penalized (set to target average).
Highly Unrealistic Models

This is one model, *not* a superposition.

Keep your friends *very* close.

Miro? Picasso? Kandinsky?
Knots in protein models

- Several models with local knots were identified by visual inspection.
- Since three targets in CASP7 contain knots (e.g. T0320, T0332, T0378), we did not explicitly exclude models with knots from the assessment.
Highly unrealistic models

- No of groups with at least one model flagged as unfeasible: 81

Number of models flagged as unfeasible per group:

Number of models flagged as unfeasible per target:
Numerical assessment of models in the TBM category
TBM Numerical Model Assessment

- **GDT (Global Distance Test)**
  - **GDT_TS** = \(\frac{GDT_{P1} + GDT_{P2} + GDT_{P4} + GDT_{P8}}{4}\)
  - **GDT_HA** = \(\frac{GDT_{P0.5} + GDT_{P1} + GDT_{P2} + GDT_{P4}}{4}\)
  - **GDT_TL** = \(\frac{GDT_{P0.25} + GDT_{P0.5} + GDT_{P1} + GDT_{P2}}{4}\)
  - with: \(GDT_{Pn}\) = percent of residues under distance cutoff \(<= n\) Å
  - Calculated with LGA (Local-Global Alignment) \(^{[1]}\)

- **AL_0**
  - Number of correctly aligned residues \((d < 3.8\) Å\) in the 5 Å LGA superposition of the model and experimental structure of the target.

Which cut-off values to use for GDT?

- GDT_TS (1,2,4,8) and GDT_HA (0.5,1,2,4) are highly correlated.
- Lower cutoff values are better suited to measure differences in good template based models [1].
- CASP7 TBM category has many good template based models which can hardly be distinguished based on GDT_TS
- Use GDT_HA for TBM assessment in CASP7.

Are GDT and AL_0 scores sufficient?

- Alignment accuracy AL_0 is highly correlated with GDT.
- Both AL_0 and GDT only capture global Cα positions.
- Cα atoms only represent 12% of the atoms in a protein structure.
- Neither AL_0 and GDT do account for:
  - Back-bone geometry
  - Side chain modeling
  - Local interactions such as hydrogen bonds or hydrophobic contacts
Assessing local atomic interactions

- Additional assessment criteria:
  - Atom contact map overlap (CMO)
  - Hydrogen bond conservation (HBscore)
  - Side chain conformation (see: HA-TBM)
  - Crystallographic electron density map correlation / phasing in molecular replacement (see: HA-TBM)
Hydrogen bond conservation ("HBscore")

\[
\text{HBscore} = \frac{\text{number of conserved H - bonds in model}}{\text{total number of H - bonds in target structure}}
\]
Hydrogen bond conservation ("HBscore")

\[
HBscore = \frac{\text{number of conserved H-bonds in model}}{\text{total number of H-bonds in target structure}}
\]

- Hydrogen bonds definition using HBPlus \[1\]
- Relative solvent accessibility calculation: NACCESS \[2\]
- Considered Hydrogen-bonds \[3\]:
  - Back bone – back bone hydrogen bonds
  - Hydrogen bonds involving buried side chain atoms (i.e. side chain with less than 50% relative surface exposure)

---

\[1\] McDonald IK & Thornton JM (1994), 'Satisfying Hydrogen Bonding Potential in Proteins', Journal of Molecular Biology 238:777-793

\[2\] Hubbard SJ & Thornton JM (1993), 'NACCESS', Department of Biochemistry and Molecular Biology, University College London.

\[3\] Residues with correct connectivity / topology, accounting for ambiguity in IUPAC atom nomenclature of chemically equivalent atoms (e.g. Glu OE1, OE2)
Hydrogen bond conservation ("HBscore")

Example:

Target

Model Group A
GDT_TS: 35.2
HBScore: 41.1

Model Group B
GDT_TS: 45.6
HBScore: 26.0
Statistical significance and ranking
Top 25 groups were selected for paired Student t-test head-to-head comparison of significant differences in GDT, AL0 and HBscore on common targets.

Negative Z-scores were set to zero. Unfeasible models were penalized (raw values set to target average / Z-score to zero). [1]

[1] Penalization of unfeasible models did not affect the ranking of the top groups.
Statistical significance and ranking

- For each of the top 25 groups selected on combined GDT / AL0 Z-scores, all against all **pair wise Student t-tests** on raw scores for common target domains were calculated.

- Ranking based on significant differences in t-tests \( p < 0.05 \) for
  - GDT_HA
  - AL0
  - HB-Score

- Direct head-to-head comparison of all 25 groups:
  
  \# Counting the number of statistically significant wins against all other groups on common targets.
Head to head comparison (GDT, AL0, HB)

- Zhang 24
- Baker 20
- Zhang-Server 25
- 556 LEE
- 125 TASSER
- 675 Fams-ace

Groups

% significant wins

0 10 20 30 40 50 60 70 80

GDT_HA
AL0
HB-Score
Total Sum

CASP7, Asilomar 28. November 2006, Torsten Schwede (23)
GDT-TS Z-score by target difficulty

All TBM targets

Easy TBM targets

Medium TBM targets

Hard TBM targets

Groups

GDT_Ts Z-score

Groups

GDT_Ts Z-score

Groups

GDT_Ts Z-score

Groups
Prediction Methods
by best scoring groups.
Methods of best scoring groups

24 / 25 **Zhang / Zhang-Server: Protein structure prediction by iterative TASSER simulations**

*Server prediction by iterative TASSER simulations in CASP7*

Yang Zhang, Center for Bioinformatics and Department of Molecular Bioscience, University of Kansas, Lawrence, KS 66047, USA.

20 **Baker: Template-based Structure Prediction in CASP7 by Rosetta and Rosetta@home**

B. Qian, V. Sraman, S. Khare, R. Das, W. Sheffler, D. Chivian, D. Kim, L. Malmstrom, A. Wollacott, D. Baker, University of Washington, Seattle WA, USA.

125 **TASSER: TASSER for protein structure prediction in CASP7**

H. Zhou, S.Y. Lee, S. Pandit, H. Chen, J. Borreguero, and J. Skolnick, Center for the Study of Systems Biology, Georgia Institute of Technology, Atlanta, USA.

556 **LEE: Template based Modeling based on Global Optimization**

K Joo, J Lee, S Lee, J Seo, SJ Lee, J Lee, Korea Institute for Advanced Study, Hanyang University, Seoul National University, Suwon University, Korea.

675 **FAMS-ACE: Model selection from server results using original threading (3D1D) program and consensus**

M. Iwadate, K. Kanou, G. Terashi, D. Takaya, K. Ohta, A. Hosoi, M Takeda-Shitaka and H. Umeyama, Department of Biomolecular Design, School of Pharmaceutical Sciences, Kitasato University, Tokyo, Japan.
General Questions for all TBM groups

- **Server vs. human**: What role did servers & automated procedures play in your approach? What role did manual intervention play? Did you improve over automated predictions?
- **Progress since CASP6**: Which algorithmic developments did you implement since CASP6 to improve your predictions in CASP7?
- **Model building**: How did you include multiple templates in modeling?
- **Refinement**: Did you apply some kind of model refinement protocol? Which methods did you use? How much attention did you give to modeling local interactions, e.g. side chain conformations?
- **Computational effort**: How computationally expensive is your approach? How much CPU time did you use per model in CASP7?
Examples of outstanding predictions by individual groups. [1]

Prediction Success

GDT_TS vs. Target Difficulty

Target difficulty ranking

GDT_TS

- Best GDT-TS model
- Mean GDT_TS of 25 best models

T0283 (20)
T0316_D3 (677)
T0321_D1 (415)
T0299_D2 (20)
T0354 (276)
T0356_D2 (212)
Prediction Success: T0356_D2

<table>
<thead>
<tr>
<th>Target</th>
<th>212/213 (HHpred1, HHpred2)</th>
<th>214 (BayesHH)</th>
<th>418 (HHpred3)</th>
<th>92 (MaOPUS-Server)</th>
<th>675 (FAMS-ace)</th>
<th>245 (UNI-EID_expm)</th>
<th>383 (UNI-EID_bnmx)</th>
<th>AL_0</th>
</tr>
</thead>
</table>

Servers in blue

Graph showing distance Cx/CA vs. percent of residues (CA).
Prediction Success: T0321_D1

Target
T0321_D1
Best template 1f9c

Template 2mnr Identified by group 415 (SP4, server)
Prediction Success: T0283

T0283 Target

GDT_TS

20 (Baker) AL0: 77.32
13 (Jones-UCL) AL0: 62.89

Target T0283

CA-CA difference, Å

Residue
Prediction Success: T0283

T0283 Target

T0283 Template: 2b2j

T0283 Side Chains

T0283 Group 20

T0283 Template: 2b2j

20 (Baker) AL0: 77.32

13 (Jones-UCL) AL0: 62.89
Predictions vs. Best Single Template.
Predictors vs. Best Single Template

Plot showing GDT-TS Z-score for different groups:
- Best template average
- Group 24
- Groups 20, 25, 125
- Group 675
Improvement over Single Templates

# number of models better than a single best template.

# number of models worse than a single best template.
Improvement over Single Templates

# number of models better than a single best template.

# number of models worse than a single best template.
Template vs. Predictors (by Target Difficulty)

Easiest targets

Mean GDT-TS per group

Groups

Template (red)
Best Group (24)
Template vs. Predictors (by Target Difficulty)

- Easiest targets
- Medium targets
- Hardest targets

Mean GDT-TS

Template (red)
Best Group (24)
Improvement over Single Templates

Difference in GDT_TS between best single template and models by the top 6 groups.

CASP7 Targets sorted by relative difficulty
Comparison CASP 6 vs. CASP 7
TBM target difficulty ranks for CASP 6 and 7

Difficulty = \{ \text{Rank} (\% \text{ID}) \\
+ \text{Rank} (\text{LGA}_S) \\
+ \text{Rank} (\text{STR}_\text{ALN}) \} / 3

Difference in average relative target difficulty of the TBM subset of CASP 6 (99.1) and CASP 7 (90.4) is statistically not significant. (Wilcoxon Rank Sum Test, p = 0.31)
Comparison CASP 6 & 7

- Blue squares represent CASP 7 Targets.
- Red triangles represent CASP 6 Targets.
- Blue line represents Poly. (CASP 7 Targets).
- Red line represents Poly. (CASP 6 Targets).

Best model AL0 for each target.
Comparison CASP 6 & 7

Mean of best 25 models AL0 for each target.
Comparison CASP 6 & 7

Mean of best 25 models GDT-TS for each target.

Target difficulty ranking

CASP 7 Targets □ CASP 6 Targets ▲ Poly. (CASP 6 Targets) — Poly. (CASP 7 Targets)
Mean of best 5 models GDT-TS for each target.
Comparison CASP 6 & 7

Best models GDT-TS for each target.
Model Confidence ("Model B-factor")

"... a model must be wrong, in some respects -- else it would be the thing itself. The trick is to see ... where it is right."

Henry A. Bent
Do we know when our models are inaccurate?

- Assessed TBM predictions of 60 Groups (32 %) which had
  - more than 10 TS predictions with “B-Factors” submitted, with
  - at least 2 different numerical values in the B-factor column.

- Assessment Approach:
  1. Linear correlation
  2. ROC Curves

- Distance measure: distance of C_α_-atoms between target structure and superposed model (LGA, sda)

- Random model T000: "Model B-factor" is randomly chosen from a list of typical model-target distances.
Model Confidence Estimates ("Model B-factor")

- Logarithmic scatter plot: Predicted model error ("B-Factor" vs. real model error for individual groups)

000 random model

046

\[
\log \left( \text{Ca dist.} + 1 \right) \\
\log( \text{"B-factor"} + 1 )
\]
Model Confidence Estimates ("Model B-factor")

- Linear correlation is sensitive to outliers:
  
  $r = 0.65$
  
  $r = 0.32$
“B-Factor” outliers T0342

PDB-Format ≠ CASP-Format?

<table>
<thead>
<tr>
<th>PFRMAT</th>
<th>TS</th>
</tr>
</thead>
<tbody>
<tr>
<td>TARGET</td>
<td>T0342</td>
</tr>
<tr>
<td>MODEL</td>
<td>1</td>
</tr>
<tr>
<td>PARENT</td>
<td>1G0Q_A 1VKB_A 1XHS_A</td>
</tr>
</tbody>
</table>

ATOM 1105  N  SER  74  33.017  21.048  10.027  1.00 88.72
ATOM 1106  CA  SER  74  32.858  21.286  11.430  1.00 88.72
ATOM 1107  C  SER  74  32.416  20.019  12.077  1.00 88.72
ATOM 1108  O  SER  74  31.420  19.408  11.691  1.00 88.72
ATOM 1109  CB  SER  74  31.868  22.411  11.660  1.00 88.72
ATOM 1110  OG  SER  74  32.344  23.638  11.177  1.00 88.72
ATOM 1111  H  SER  74  32.338  20.500  9.519  1.00 88.72
ATOM 1112  HG  SER  74  31.690  24.321  11.341  1.00 88.72
ATOM 1113  HA  SER  74  33.755  21.686  11.906  1.00 88.72
ATOM 1114  1HB  SER  74  30.938  22.166  11.149  1.00 88.72
ATOM 1115  2HB  SER  74  31.682  22.500  12.729  1.00 88.72
ATOM 1116  N  PRO  75  33.153  19.598  13.060  1.0046657.84
ATOM 1117  CA  PRO  75  32.716  18.436  13.770  1.0046657.84
ATOM 1118  C  PRO  75  31.580  18.860  14.635  1.0046657.84
ATOM 1119  O  PRO  75  31.627  19.964  15.171  1.0046657.84
ATOM 1120  CB  PRO  75  33.961  17.983  14.551  1.0046657.84
ATOM 1121  CG  PRO  75  34.766  19.236  14.759  1.0046657.84
ATOM 1122  CD  PRO  75  34.546  19.985  13.376  1.0046657.84
ATOM 1123  1HD  PRO  75  34.507  20.927  13.519  1.0046657.84
ATOM 1124  2HD  PRO  75  35.034  19.671  12.620  1.0046657.84
ATOM 1125  1HG  PRO  75  34.396  19.695  15.368  1.0046657.84
ATOM 1126  2HG  PRO  75  35.663  19.026  14.588  1.0046657.84
ATOM 1127  1HB  PRO  75  33.587  17.773  15.359  1.0046657.84
ATOM 1128  2HB  PRO  75  34.367  17.506  13.954  1.0046657.84
ATOM 1129  HA  PRO  75  32.424  17.732  13.128  1.0046657.84

[...]

CASP7, Asilomar 28. November 2006, Torsten Schwede (52)
ROC Curves

- **Model-Error Classification**
  - Residue Cα error $\geq$ 3.8 Å: incorrect modeled residue
  - Residue Cα error $< 3.8$ Å: correct modeled residue

- **Enrichment of correctly identified model errors**
  - "Model B-factor" values re-ranked between 0 and 1.
  - Plot: Enrichment of correctly identified model errors
  - Ranking by “area under the curve”
Model Confidence Estimates ("Model B-factor")

ROC Curves (3.8 Å)
Performance of Model Confidence Estimates
Best Groups in “B-Factor” Assessment:

<table>
<thead>
<tr>
<th>Groups</th>
<th>ROC (AUC)</th>
<th>correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>050 SBC</td>
<td>0.858</td>
<td>0.528</td>
</tr>
<tr>
<td>046 Pcons6</td>
<td>0.844</td>
<td>0.616</td>
</tr>
<tr>
<td>047 Pmodeller6</td>
<td>0.832</td>
<td>0.573</td>
</tr>
<tr>
<td>005 luethy</td>
<td>0.771</td>
<td>0.526</td>
</tr>
<tr>
<td>214 BayesHH</td>
<td>0.765</td>
<td>0.434</td>
</tr>
<tr>
<td>275 beautshot</td>
<td>0.761</td>
<td>0.446</td>
</tr>
<tr>
<td>038 GeneSilico</td>
<td>0.753</td>
<td>0.287</td>
</tr>
<tr>
<td>004 ROBETTA</td>
<td>0.742</td>
<td>0.492</td>
</tr>
</tbody>
</table>
“Model B-Factor” Conclusions

- Several groups have provided realistic model accuracy estimates in CASP7. Best results by group 50 (SBC).

- The best groups were using an absolute metric in Ångstrom (as requested by the CASP format). However, many groups are using their own metric.

- Unfortunately, only 32% of the groups have provided confidence values for their predictions. Therefore, it was not possible to use confidence factors in the structural assessment.
Acknowledgement

A big “Thank you!”

to all predictors for their
tremendous job in building more
than 15’000 models in the CASP7
TBM category!