Target Classification in the 14th Round of the Critical Assessment of Protein Structure Prediction (CASP14)

Lisa Kinch, Andriy Kryshtafovych and Nick Grishin
CASP14 Target Classification during COVID-19

What used to be

Start of Registration

Los Angeles, March 30 2020

Registration for CASP14 opened, March 9, 2020

Can we get enough targets?
CASP14 Domain Definition and Evaluation Units

What used to be

- Release Sequences for Targets
- Collect predictions
- Define Domain Bounds:
  - DomainParser (Prediction Center)
  - Sequence continuity vs Structure compactness
  - Template Domains (ECOD)
- Trim predictions
- Redefine bounds
- Calculate Scores on domains:
  - Top20 Server Performance
  - Grishin Plots, New Templates
- Informed decision

Domains

Evaluation Units (EVUs)
CASP14 Pre-Evaluation Domain Definition

What used to be

- Release Sequences for Targets
- Collect predictions
- Define Domain Bounds:
  - DomainParser (Prediction Center)
  - Sequence continuity vs Structure compactness
  - Template Domains (ECOD)
- Trim predictions
- Redefine bounds
- Calculate Scores on domains:
  - Top20 Server Performance
  - Grishin Plots, New Templates
- Trimmed predictions
- Split into EVUs
- Merge Domains

CASP14 adaptation

- T1044 Define Domain Bounds:
  - Suggestion from Experimentalist
  - Sequence continuity vs Structure compactness
  - Template Information

Bypasses the rigorous process for defining domains
CASP14 Pre-Evaluation Leak of Information

**T1044:** Phage DNA-dependent RNA polymerase

2xRIFT-related Active site + “connector”

Preprint Statistics (Andriy): CASP boosts Interest 5-fold in 1 week
**T1044 Pre-Evaluation** split into 9 Targets

**T1044**: Phage DNA-dependent RNA polymerase

*exclude* 2xRIFT-related Active site + "connector"

Allows a Comparable Number as in previous CASPs
**Template Information was Lacking**

**T1044**: Phage DNA-dependent RNA polymerase

Evolutionary Relationships
- **New Fold** (4)
- **Topology-level** (4)
- **Distant Homolog** (1)

Lack of Templates and extensive domain interactions mean Domains might not be independent folding units
T1027: Gaussia luciferase
• NMR structure with high flexibility
• Loose ensemble
• 5 disulfide bond pairs

- Keep overlapping parts of the structure
- Trim last disulfide pair

Keep trimmed model1 as the T1027 EVU
Domains Have Many Different Definitions

What is a Domain?

- Compact, globular substructures that have more interactions within them than with the rest of the structure

![Protein kinase-like Domain](Protein kinase-like Domain)

**Alpha-helical bundle**
Domains are More than Compact Substructures

What is a Domain?

• Compact, globular substructures that have more interactions within them than with the rest of the structure

• Conserved, Independent folding unit that can exist in multiple contexts, i.e. serve as building blocks of evolution

• Evolution tends to preserve sequence continuity in domains

Simple domain organization

T1053
What is a Domain?

- Compact, globular substructures that have more interactions within them than with the rest of the structure
- Conserved, Independent folding unit that can exist in multiple contexts i.e. serve as building blocks of evolution
- Evolution tends to preserve sequence continuity in domains
- Evolutionary Classification of Protein Domains (ECOD) database was an essential resource for defining domains: prodata.swmed.edu/ecod/ (thanks Dustin!)

ECOD PMID: 25474468
Domains = Evaluation Units

• Using split domains as EVUs are required when templates have known conformation changes (example to follow)

• Using split domains as EVUs are required when they have different difficulty levels (perhaps not in the future)

For CASP14 we tried to keep domains together; If not, we evaluated domain interactions in a separate assessment.
Domains = Evaluation Units

• Using split domains as EVUs are required when templates have known conformation changes (example to follow)

• Using split domains as EVUs are required when they have different difficulty levels

• Decisions to split or merge are based on group performance: traditionally evaluated using “Grishin Plots”
Grishin Plots Inform Decisions to Split Targets

T1053 “Grishin Plot”

Performance Increases for Split Domains

Combined Domain GDT_TS

Split Domains Weighted sum GDT_TS

T1053 = 2 EVUs

T1053-D1 T1053-D2

N-C
Merging Target Domains as Evaluation Units

No Need to Split when Good Templates Exist

Domain in virus attachment proteins

Pectin lyase-like

ECOD template 6f7dA = 2 domains

T1052
Merging Target Domains as Evaluation Units

No Need to Split when Good Templates Exist

T1052-D1 = 1 EVU?

ECOD template 6f7dA = 2 domains

Grishin Plot

Confirmed!
T1061: *E.coli* phage tail
- Complex domain organization

Some Domain Definitions are Difficult
Some Domain Definitions are Difficult

Automatic Domain Parser is non continuous

**T1061: E.coli phage tail**
- Complex domain organization
- Domain parser and Ddomain split differently (4 vs 5)
Some Domain Definitions are Difficult

T1061: *E. coli* phage tail
- Complex domain organization
- Domain parser and Ddomain split differently (4 vs 5)
- Grishin Plot has multiple clouds
**T1061: E.coli phage tail**

- Complex domain organization
- Domain parser and Ddomain split differently (4 vs 5)
- Grishin Plots have multiple clouds
- Templates for blue and red domains

Homologous Templates Suggest Domain Bounds

- 3cddF Template
  - 4 domains: RIFT-related, N0 domain, insert, and RIFT-related *but 1EVU*
- 1ten Top LGA_S Template
  - Immunoglobulin-related
Topoology-level Insert is More Difficult: Suggests a Split

= different difficulty level
So 3 EVUs

T1061: E.coli phage tail
• Complex domain organization
• Domain parser and Ddomain split differently (4 vs 5)
• Grishin Plots have multiple clouds
• Templates for blue and red domains

2yc2 Top LGA_S Template
Intraflagella Transport Protein 25
jelly-roll

2frg Top Dali Template
human TLT1
Immunoglobulin-like β-sandwich

N- -CD1 D1D2 D3

D2

Beta-sandwich
Splits for Targets that Change Conformation

T1024 MFS transporter

Outward Facing Template 3wdo

Inward Facing Template 4j05

T1024 Grishin Plot

Shifts = Conformations Change?
Splits for Targets that Change Conformation

4 Similar Targets: T1024, T1050, T1100, T1101

427_1

427_3

Outward Facing Template 3wdo

Inward Facing Template 4j05

Split Domains GDT-TS Weighted Sum

Shifts = Conformations Change?

T1024 Grishin Plot

Combined Domains GDT-TS
CASP14 Domains and EVUs in Numbers

- 67 Targets: 9 pre-split from 1 structure, 11 from 6 interacting structures
- 48 Single EVU Targets
  - 6 have 18 merged ECOD domains
- 19 Split Domain Targets (48 EVUs)
  - 4 have 10 merged ECOD domains

=96 Targets for classification into Topology-level (FM) and High Accuracy-level (TBM)

Evolutionary Relationships to known Templates help Classification
Evolution-Based Classification of CASP14 EVUs

67 Targets: 9 pre-split from 1 structure, 11 from 6 interacting structures

48 Single EVU Targets
6 have 18 merged ECOD domains

19 Split Domain Targets (48 EVUs)
4 have 10 merged ECOD domains

ECOD Classification based on distance to template:

<table>
<thead>
<tr>
<th>Class</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family (24 EVU)</td>
<td>Template is in the same cdd</td>
</tr>
<tr>
<td>H-group (50 EVU)</td>
<td>Template is homologous</td>
</tr>
<tr>
<td>X-group (12 EVU)</td>
<td>Topological similarities</td>
</tr>
<tr>
<td>New (10 EVU)</td>
<td>Unique combination of SSEs</td>
</tr>
</tbody>
</table>

- 13 Immuno-globulin-like
- 9 Cradle-loop barrels
- 4 α/β-hammer-heads
- a+b three-layered sandwiches
- alpha arrays
- a+b two layers
- beta sandwiches
- extended segments
- a+b complex topology
- a+b duplicates or obligate multimers
- alpha duplicates or obligate multimers
- alpha superhelices
- alpha complex topology
- beta complex topology
- beta duplicates or obligate multimers
- a+b three layers
### ECOD Architectures: CASP14 compared to CASP13

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Observed (CASP13 or CASP14 frequencies)</th>
<th>Expected (ECOD Family frequencies)</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta complex topology</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>a+b complex topology</td>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td>alpha complex topology</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>alpha duplicates or obligate multimers</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>beta duplicates or obligate multimers</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>beta barrels</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>alpha superhelices</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>beta sandwiches</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>a+b duplicates or obligate multimers</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>alpha arrays</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>a+b two layers</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>alpha bundles</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>a+b three layered sandwiches</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>a+b three layers</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>extended segments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a/b barrels</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a+b four layers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>few secondary structure elements</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mixed a+b and a/b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>beta meanders</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The chart compares the observed frequencies of different ECOD architectures in CASP13 and CASP14 to the expected frequencies based on ECOD family frequencies.
### ECOD Relationships: CASP14 compared to CASP13

#### Suggests Increased Difficulty

<table>
<thead>
<tr>
<th>Structure Type</th>
<th>CASP14 Count</th>
<th>CASP13 Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta complex topology</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>a+b complex topology</td>
<td>18</td>
<td>4</td>
</tr>
<tr>
<td>alpha complex topology</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>alpha duplicates or obligate multimers</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>beta duplicates or obligate multimers</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>beta barrels</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>alpha superhelices</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>beta sandwiches</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>a+b duplicates or obligate multimers</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>alpha arrays</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>a+b two layers</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>alpha bundles</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>a/b three-layered sandwiches</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>a+b three layers</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>extended segments</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>a/b barrels</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>a+b four layers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>few secondary structure elements</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mixed a+b and a/b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>beta meanders</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **CASP14** compared to **CASP13** with a trend suggesting increased difficulty.
Traditional CASP Classification Plot: Scatter is Broad

**EVU Classification Scores**

Hhscore = HHprobability x Coverage for Chosen Template

Use higher HHscore from 2 methods: Uniprot100 or PDB70 for query profile

Select Rank1 template unless

- Max HHscore > for alternate homolog
- Lower rank homolog replaces analog

Top LGA_S from homolog or analogous fragment

CASP13
Traditional CASP Classification Plot: Scatter is Broad

**EVU Classification Scores**

\[ \text{Hhscore} = \text{HHprobability} \times \text{Coverage for Chosen Template} \]

- Use higher Hhscore from 2 methods: Uniprot100 or PDB70 for query profile
- Select Rank1 template unless:
  - Max Hhscore > for alternate homolog
  - Lower rank homolog replaces analog
- Top LGA_S from homolog or analogous fragment

Where to draw the boundaries for FM and TBM?
What Contributes to Broadened Scatter?

Sequence Component

Traditional Plot

Structure Component

CASP13

CASP14

Blurred Boundary

Analogs

Homologs

Performance (Top 20 Server Avg GDT)

Performance (Top 20 Server Avg GDT_Ts)

Top Template (LGA_S)

Top Template (LGA_S)
Cluster Data to Help Confirm Classification Bounds

Heatmap with hierarchical Clustering

Data Matrix

Features

Principal Component Analysis (PCA)

PMID: 25969447
PCA Plot of Targets Roughly Separates Classes

Scores Used:
- HHscore
- %parentTBM
- Neff%max
- performance
- TopLGA
- Dali%self
- DaliCvg

Data Preprocessing:
- No scaling, rows centered

PCA Method:
- SVD with Imputation

Prediction ellipses:
- Probability 0.95

Feature = CASP Classification
PCA Plot of Targets Roughly Separates ECOD Groups

Feature = ECOD Classification

T1096-D1
Phage RNA Pol Subunit

3les RNA Pol
Sigma Factor
No scaling is applied to rows. Imputation is used for missing value estimation. Rows are clustered using correlation distance and Ward linkage. Columns are clustered using Euclidean distance and Ward linkage. 7 rows, 96 columns.
No scaling is applied to rows. Imputation is used for missing value estimation. Rows are clustered using correlation distance and Ward linkage. Columns are clustered using Euclidean distance and Ward linkage. 7 rows, 96 columns.
Heatmap Clusters Targets by Classes

*No scaling is applied to rows.* Imputation is used for missing value estimation. Rows are clustered using correlation distance and Ward linkage. Columns are clustered using Euclidean distance and Ward linkage. 7 rows, 96 columns.

Target Clustering

- **T1052-D2**
- **T1052-D3**
- **T1055**

**Sec63 N-domain-like**

**HEH Motif**

**MergedDomains:** 

**HHscore for HEH**
Domains at the edge: i.e. near the boundary in the traditional classification

Most domains were classified by the traditional scatter (to be consistent with CASP13)

T1055, T1052-D2 and T1052-D3 cluster differently by heatmaps, but are classified by the scatter
Domains at the edge: i.e. near the boundary in the traditional classification

Most domains were classified by the traditional scatter (to be consistent with CASP13)

T1055, T1052-D2 and T1052-D3 cluster differently by heatmaps, but are classified by the scatter

T101-2 and T1065s2 cluster differently by the scatter, but are classified by the heatmap groups
Traditional CASP Classification Plot: Outliers

Domains at the edge: i.e. near the boundary in the traditional classification

Most domains were classified by the traditional scatter (to be consistent with CASP13)

T1055, T1052-D2 and T1052-D3 cluster differently by heatmaps, but are classified by the scatter

T101-2 and T1065s2 cluster differently by the scatter, but are classified by the heatmap groups

T1065s2 New fold with increased Performance

T1101-D2 Family level with shifted LGA
Domains at the edge: i.e. near the boundary in the traditional classification

Most domains were classified by the traditional scatter (to be consistent with CASP13)

T1055, T1052-D2 and T1052-D3 cluster differently by heatmaps, but are classified by the scatter

T101-2 and T1065s2 cluster differently by the scatter, but are classified by the heatmap groups
Thank You!

Collaborators
Nick Grishin (UTSW)
Dustin Schaeffer (UTSW)
Jimin Pei (UTSW)
Andriy Kryshtafovych (Prediction Center)

CASP Assessors
Andrei Lupas (High Accuracy Models)
Alfonso Valencia (Contacts)
Daniel Rigden (Refinement)
Ezgi Karaca (Assembly)
Chaok Seok (Model Accuracy)
Sandor Vajda (Function)

CASP Organizing Committee
John Moult, CASP chair and founder; IBBR, University of Maryland, USA
Krzysztof Fidelis, founder, University of California, Davis, USA
Andriy Kryshtafovych, University of California, Davis, USA
Torsten Schwede, University of Basel, Switzerland
Maya Topf, Birkbeck, University of London, UK