Domain definition and classification

BK Lee
National Cancer Institute
National Institutes of Health
Acknowledgement

Our team
Chin-Hsien (Emily) Tai
Todd Taylor
Hongjun Bai

UCSF Chimera team
Conrad Huang
Elaine Meng
Eric Petterson
Thomas Ferrin
Tom Goddard

CASP10 organizers
Andriy Kryshtafovyich
Anna Tramontano
John Moult
Krzysztof Fidelis
Torsten Schwede

CASP10 assessors
Gaetano Montelione’s team
David Jones’s team

NIGMS and EMBO
Structural biologists and Prediction groups
## Number of targets

- **Total number of structures**: 114  
  
  (644 to 758 except 748)

- **Cancelled**: 18  
  
  - No structure: 7  
  - Leaked: 9  
  - Sequence error: 1  
  - Too many residues missing: 1

- **Number after cancellations**: 96

- **Number of target domains**: 131  
  
  - TBM: 111  
  - FM: 19 (-1 server-only)  
  - TBM/FM: 1
Domain definition procedure

Generally, domains are manually defined for each target. A domain is recognized as a unit
• About 40 residues or larger and
• Physically separated from the rest of the protein or
• Have a distinctive architecture that is different from the neighboring parts of the chain.

Domain Parser2 and PDP outputs were consulted. Final decision was made after discussing with Guy’s team.
Domain definition issues

1. Two domains, one evaluation unit
   Normally, each domain is an evaluation unit (EU).
   But two domains were considered as one EU when
   • A homologous template exists that spans both domains and
   • Grishin plot is nearly diagonal.

   At least 14 cases, e.g. T0755, T0721.
Finding all templates with different inter-domain orientations

1. Find templates that have both domains of the target structure:
   - Find the set of structures (set a) that has domain D1.
   - Find the set of structures (set b) that has domain D2.
   - Find structures (set c) that are in both sets a and b.
Finding all templates with different inter-domain orientations

2. Measure the difference in angle between the two domains in the target and a template:
   • Superimpose target to the template using only D1. Call this target structure T1.
   • Superimpose target to the template using only D2. Call this target structure T2.
   • Find the angle required to best superpose T1 to T2.
T0721 templates

![Graph showing the relationship between sequence identity percentage and inter-domain angle difference in degrees. The x-axis represents the inter-domain angle difference (in degrees) ranging from 0 to 100, and the y-axis represents the sequence identity percentage ranging from 0 to 0.35. The data points are scattered across the graph, indicating variability in the relationship.]
Domain definition issues

2. Two domains arranged in an unusual manner

Each of the two domains is given a domain assignment.

In addition, the whole is given a third, D0, domain assignment. This domain is solely to evaluate the relative position and orientation of the two domains.

4 cases: T0663, T0675, T0690, T0713
T0663, T0644 and T0675
T0690 and T0713
Domain definition issues

3. **Domains with decorations**

Some proteins have a core domain + a ‘decoration’ that does not appear to be stable by itself.

Three possibilities for handling decorations:

1. Ignore or include it in the core domain(s).
2. Define it as a separate domain.
3. Define D0 as the whole protein solely to evaluate the decoration in the context of the whole.

* e.g. T0651, T0693
Proteins with a ‘decoration’
Domain definition issues

4. **Open structures**

We have two open structures, both of which were treated as one EU.
Two open structures
Domain definition issues

5. Domain-swapped dimers

Three possibilities:

– Use one chain as is.
– Define swapped part as a separate domain.
– ‘Unswap’
  
  2 cases: T0706, T0747

In both cases, we unswapped.
Domain-swapped dimers
Domain classification procedure

Generally,

TBM if structural template can be found from a sequence search,
FM if a structural template cannot be found,
Otherwise, also consult GDT scores and position-specific alignment plots of the predictions.

Structural template search programs used:

BLASTP, Pfam, TMalign
HHpred, PSIBLAST (Prediction Center)
90th percentile and maximum GDT scores
90th percentile and maximum GDT scores, magnified
Tail spike protein T0739
Thank you