TESTING TOP TESTING TOP TESTING TOP TESTING

TESTING BOTTOM TESTING BOTTOM TESTING BOT

# CASP9 Targets: domains and classifications

**Lisa N. Kinch**, **ShuoYong Shi**,
Qian Cong, Hua Cheng, Jimin Pei, Torsten Schwede & Nick V. Grishin

Howard Hughes Medical Institute and Biochemistry Department,
**University of Texas Southwestern Medical Center**,
**Dallas, Texas, USA**

**Lisa N. Kinch**
research scientist

**ShuoYong Shi**
postdoc

**Qian Cong**
graduate student

**Hua Cheng**
postdoc

**Torsten Schwede**
professor

**Jimin Pei**
research scientist

# Talk plan

- Target Overview
- Domain Definition
- Domain Classification
- CASP9 categories: TBM and FM

# Talk plan

- Target Overview
- Domain Definition
- Domain Classification
- CASP9 categories: TBM and FM

# CASP9 Target Overview

- Targets proposed:

**129** from **T0515** to **T0643**

# CASP9 Target Overview

- Targets proposed:

  **129** from **T0515** to **T0643**

- **60** targets selected for human prediction, so we have:
  server / human and server TS targets

# CASP9 Target Overview

- Targets proposed:

  **129** from **T0515** to **T0643**

- **60** targets selected for human prediction, so we have:

  server / human and server TS targets

- Targets excluded from assessment:

  **13** – for servers

  **18** – for human predictions

(**15** of them are "server", only **3** are really "human")

# CASP9 excluded targets

**For 5 targets**, it was detected that the structure was exposed in various ways:

- on the web;
- prematurely released in PDB;
- solved by a different group and released in PDB.

so human predictions were not considered, but **NONE** of these targets were actually marked as "human";

Server predictions were assessed for them.

# CASP9 excluded targets

**13** targets were canceled mostly because no experimental structure was provided in time, or it didn't correspond to sequence released for prediction.

Only **3** of these corresponded to "human" targets.

**So, as a result:**

# CASP9 assessed targets

**57** targets were assessed
for "human" predictions.

**116** targets were assessed
for "server" predictions.
These included **all** "human" targets

# Thanks

## to structural biologists who enable all this fun !

## Number of targets received from:

| | |
|---|---|
| Joint Center for Structural Genomics (JCSG) | 38 |
| Structural Genomics Consortium (SGC) | 7 |
| Midwest Center for Structural Genomics (MCSG) | 28 |
| Northeast Structural Genomics Consortium (NESG) | 39 |
| New York Structural Genomics Res. Center (NYSGXRC) | 5 |
| Non-SGI research Centers and others (Others) | 12 |

# Talk plan

- Target Overview
- Domain Definition
- Domain Classification
- CASP9 categories: TBM and FM

# Why domains?

Traditionally, CASP targets are evaluated as domains,

i.e. each target structure is **parsed into domains**,

and model quality is computed for each domain separately.

This strategy makes sense, because:

# Why domains?

**Domains can be mobile** and their relative packing can be influenced by ligand presence, crystal packing for X-ray structures, or be semi-random in NMR structures. Thus even a perfect prediction algorithm will not be able to cope with this adequately, e.g. in the absence of knowledge about the ligand presence or crystal symmetry.

# Why domains?

**Domains can be mobile** and their relative packing can be influenced by ligand presence, crystal packing for X-ray structures, or be semi-random in NMR structures. Thus even a perfect prediction algorithm will not be able to cope with this adequately, e.g. in the absence of knowledge about the ligand presence or crystal symmetry.

**Predictions may be better or worse for individual domains than for their assembly.** This happens when domains are of a different predictability, e.g.

one has a close template, but the other one does not.

Even if domains of a target are of equal prediction difficulty, it is possible that the mutual domain arrangement in the target structure, while predictable in principle, differs from the template, and thus is modeled incorrectly by predictors.

# Why domains?

**Comparison**

of the **whole-chain** evaluation
with the **domain-based** evaluation

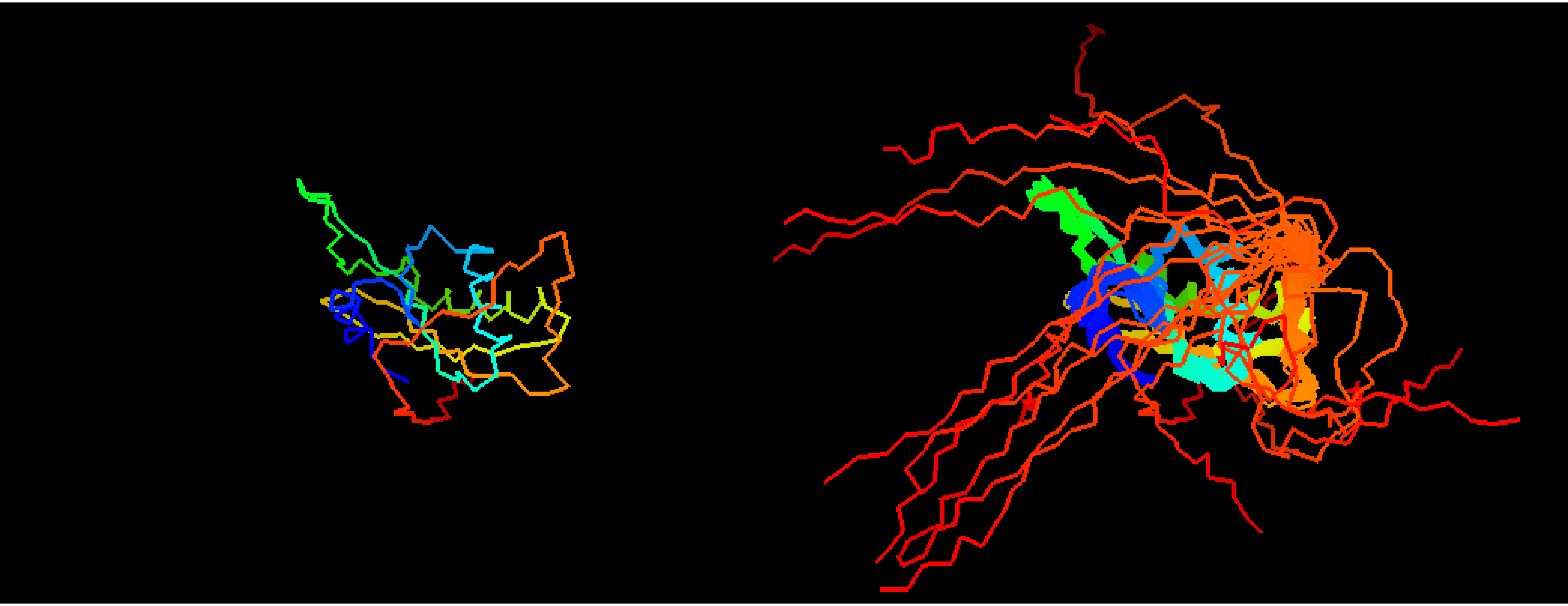
dissects the problem of 'individual domain' vs. 'domain assembly' modeling and

should aid in
development of prediction methods.

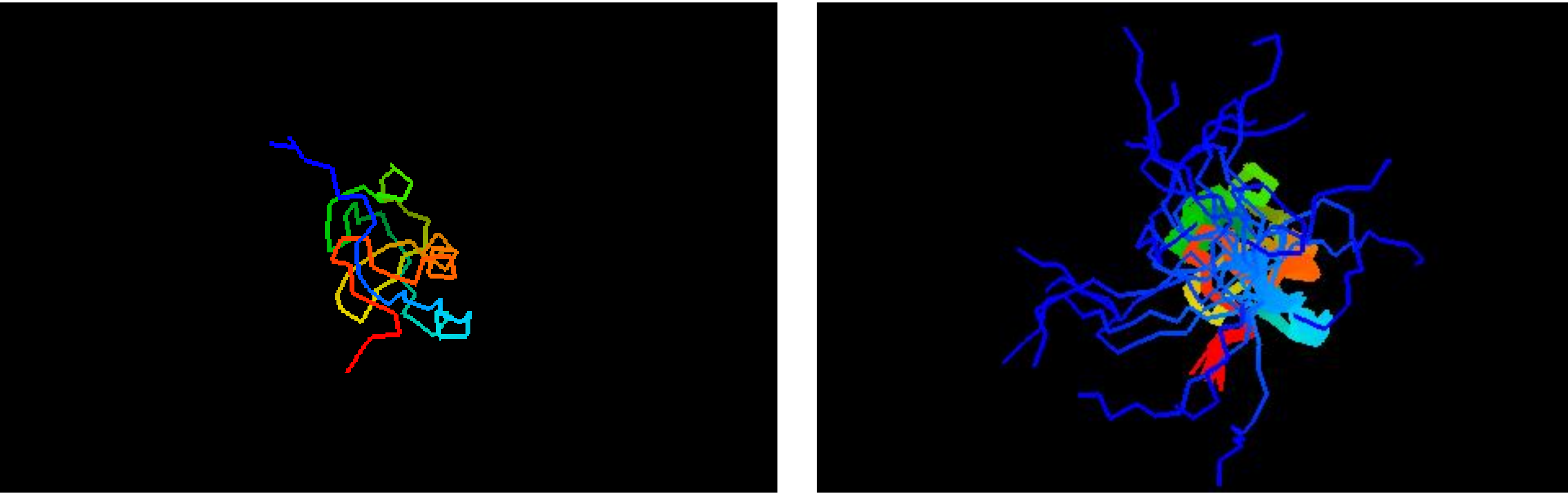
# “Whole chain” – is not the whole content of the PDB file

## NMR models: disordered regions removed!

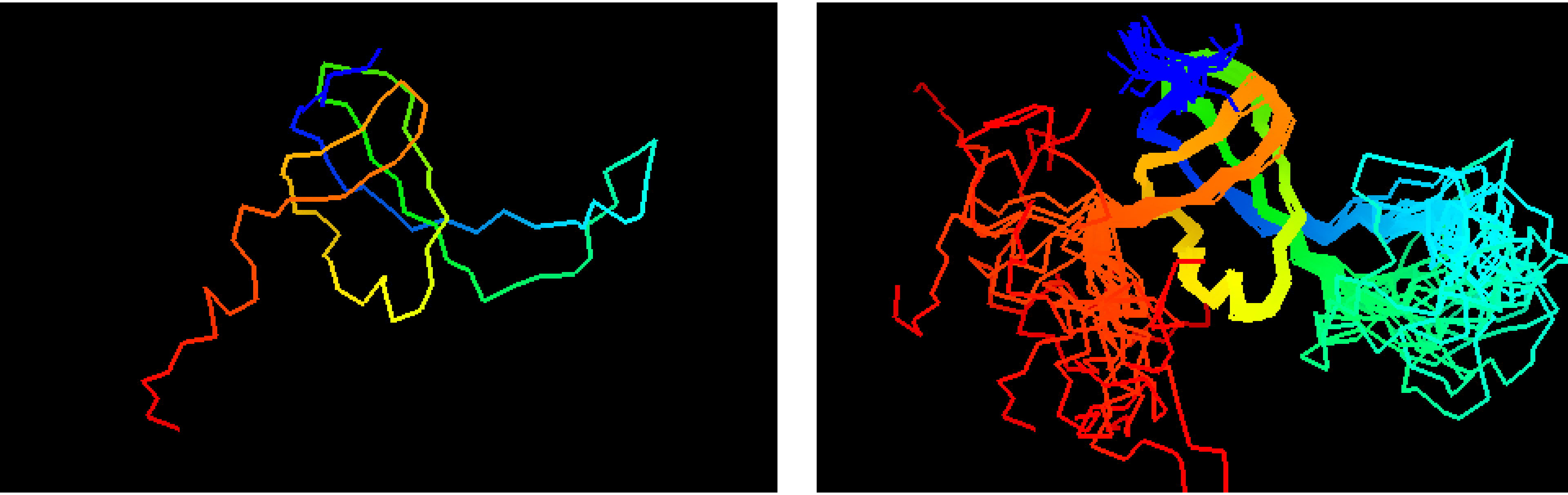
**(3.5A root mean atomic displacement in TESEUS maximum likelihood minimum RMSD superposition)**



557 NeR70A



539 RING finger



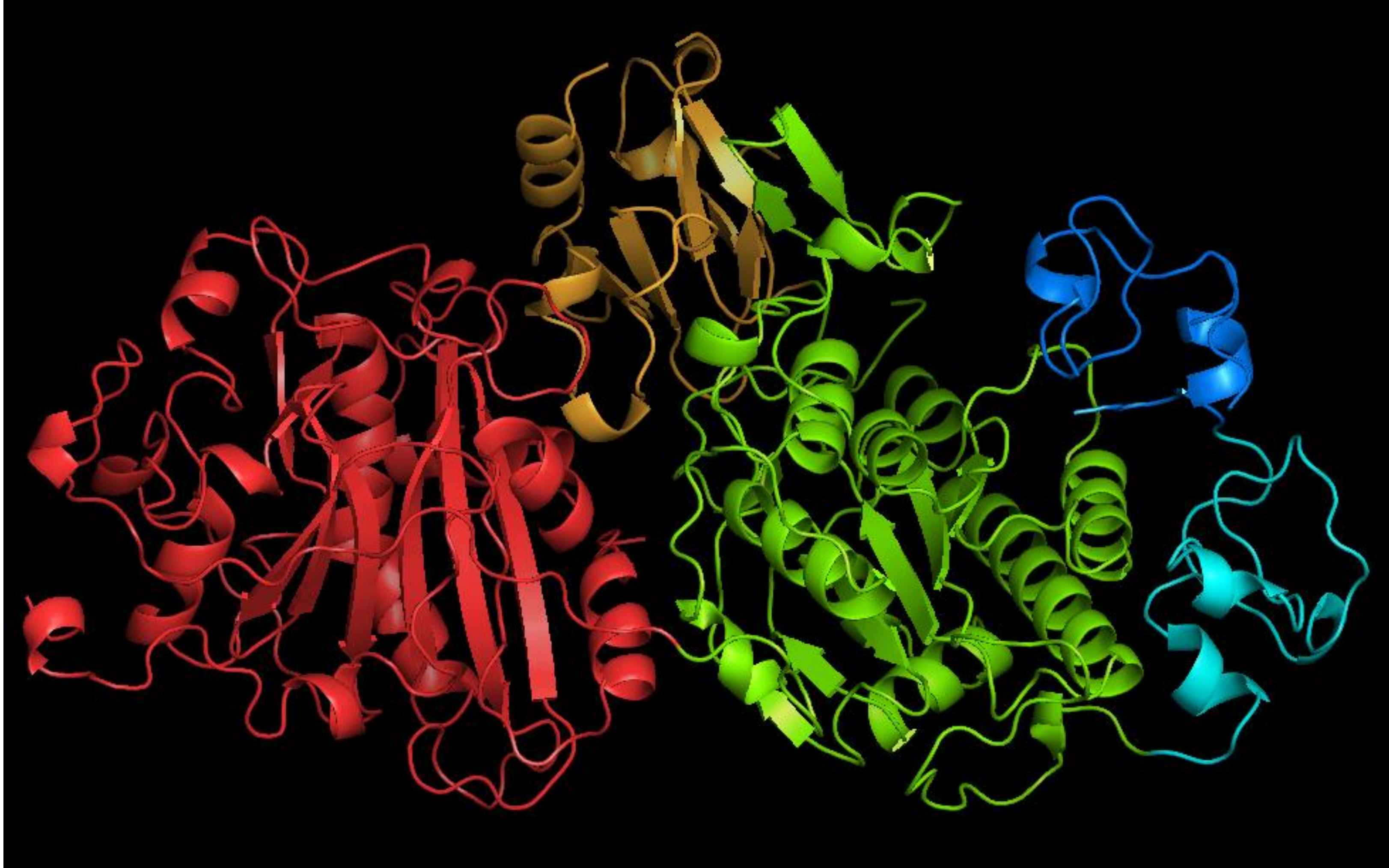
564 OB-fold

# How domains?

**Evolutionary domains**: correspond to **structurally compact evolutionary modules**

Autotaxin from rat: **T0543** consist of 5 domains

**http://prodata.swmed.edu/CASP9/evaluation/DomainDefinition.htm**



# Should we use all evolutionary domains?

**116 targets, 176 evolutionary domains, do we need that many?**

# Listen to your data!

**Cutoffs, changes, strategies should come naturally from the data you have**

# Should we use all evolutionary domains?

**116 targets, 173 evolutionary domains, do we need that many?**

Server predictions help us to reduce the number of domains:

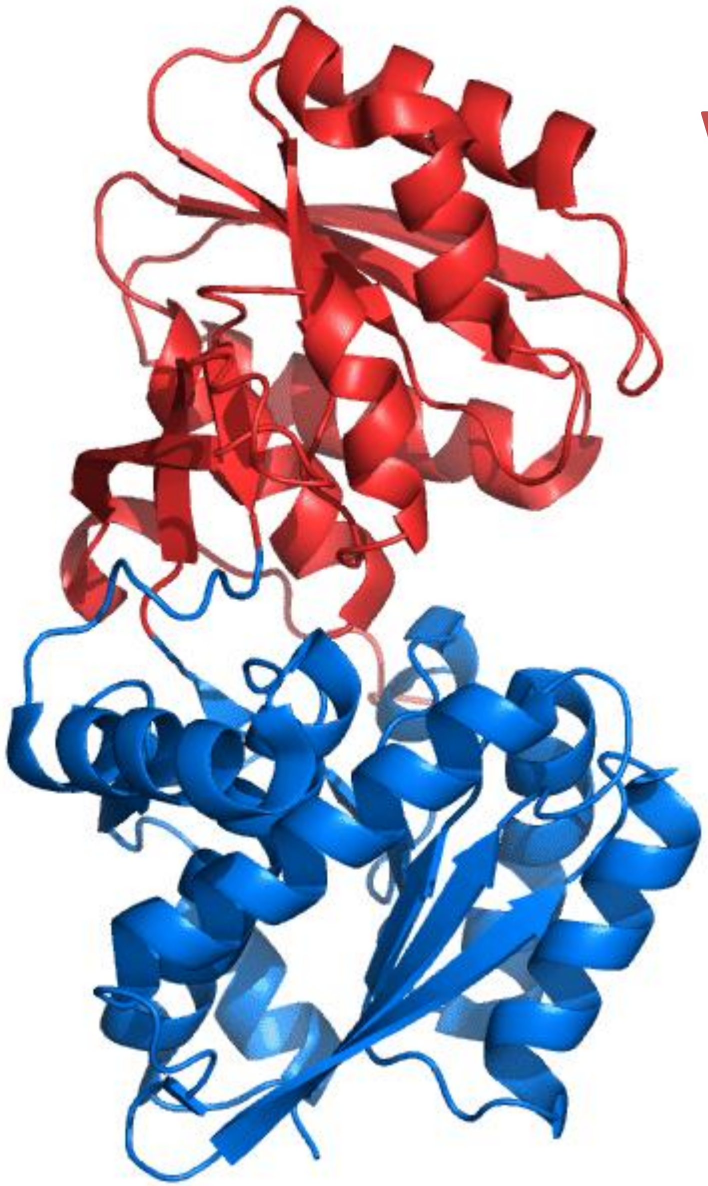
**if whole chain prediction quality is not much different from domain prediction quality, domain evaluation is not necessary.**

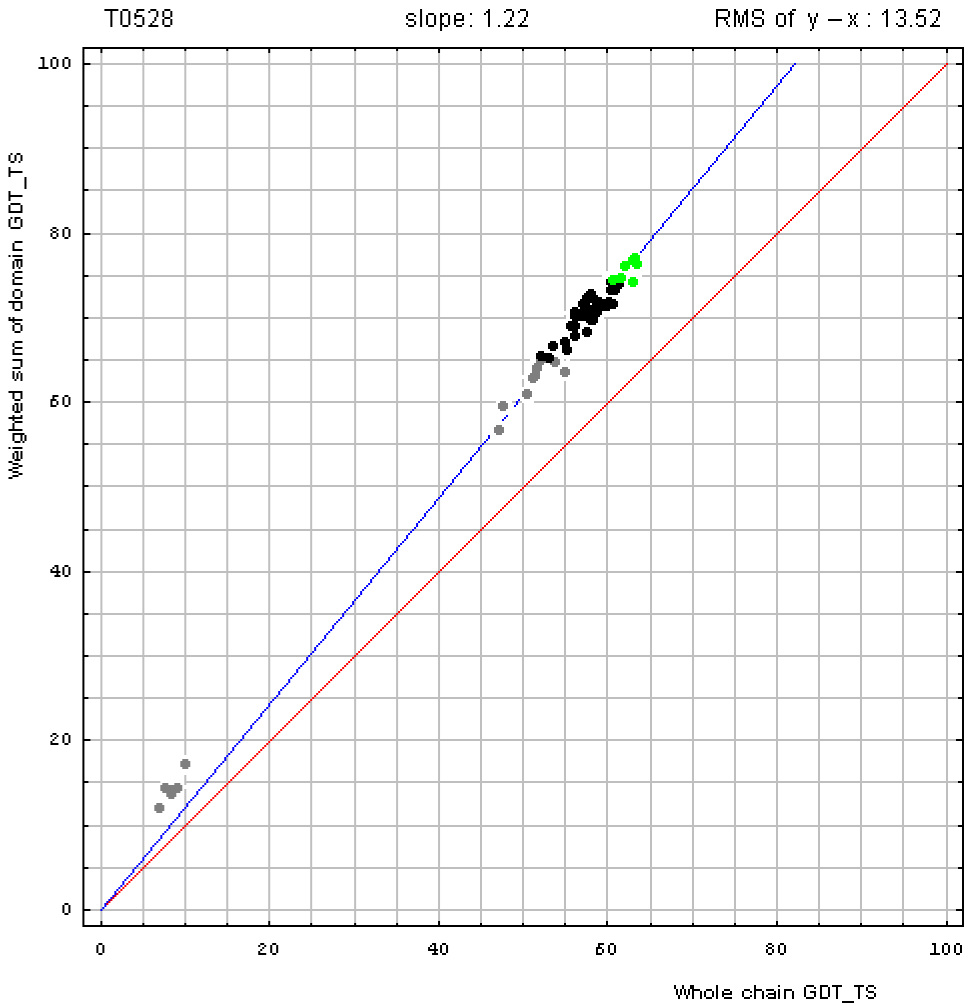
**GDT-TS(whole chain)** **vs.**

$$\frac{\sum_{i=1}^{\text{Number of domains}} \text{Length(domain i)} * \text{GDT-TS(domain i)}}{\sum_{i=1}^{\text{Number of domains}} \text{Length(domain i)}}$$

**http://prodata.swmed.edu/CASP9/evaluation/Domains.htm**

# T0528: correlation between whole chain and domain predictions





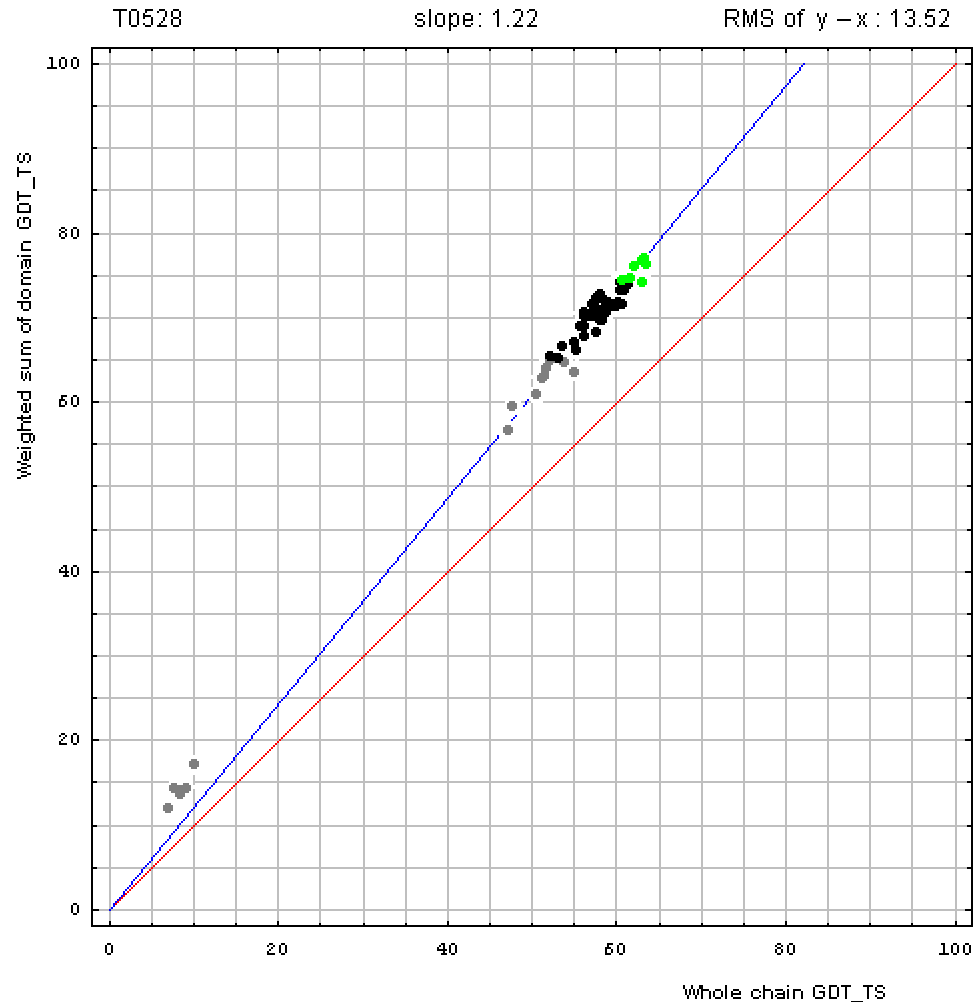

Correlation between weighted by the number of residues sum of GDT-TS scores for domain-based evaluation (y, vertical axis) and whole chain GDT-TS (x, horizontal axis).

# Two parameters to describe correlation between whole chain and domain predictions

**1.** The root mean square (RMS) difference between the weighted sum of GDT_TS on domains and GDT_TS on the whole chain (**RMS of y−x**) measures absolute GDT-TS difference.

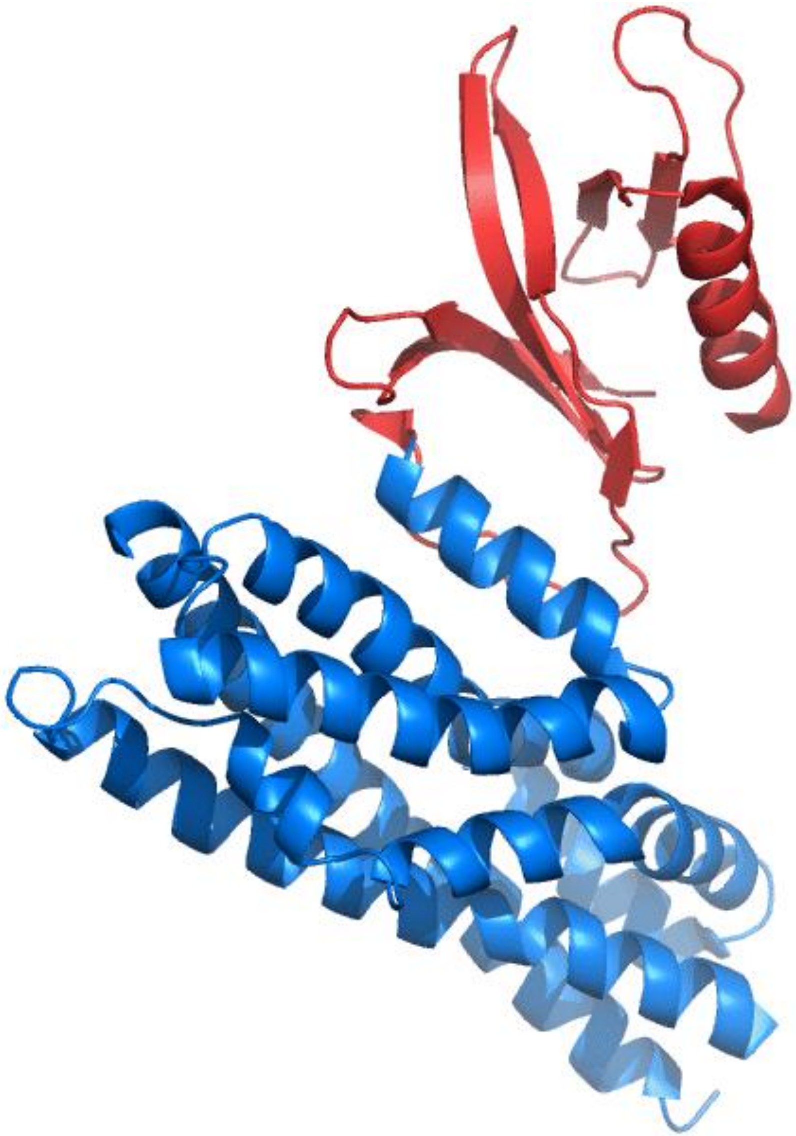
**2.** A slope of best-fit line with intercept set to 0 (**slope**) measures relative GDT-TS difference.

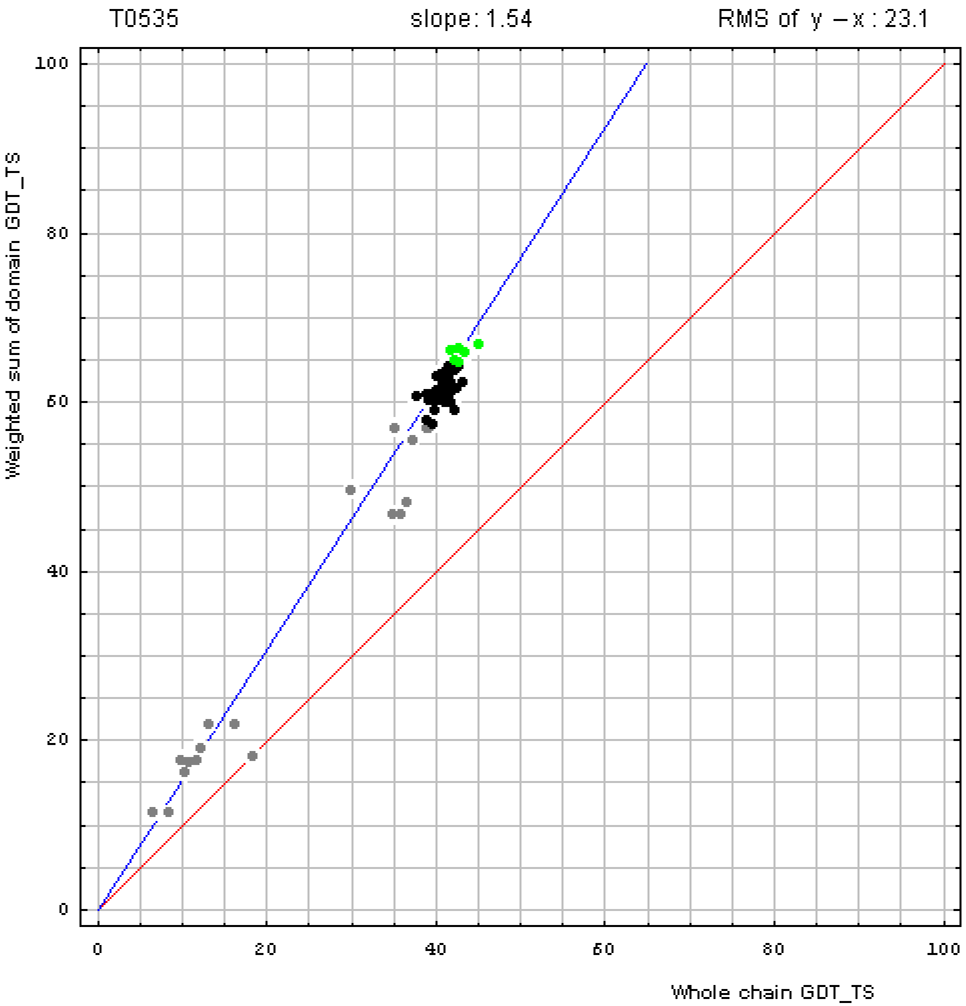
These parameters are computed on **top 10** (according to the weighted sum) **predictions**




Each point represents first server model. **Green**, **gray** and **black** points are top 10, bottom 25% and the rest of models. Blue line is the best-fit slope line (intersection 0) to the top 10 server models. Red line is the diagonal.

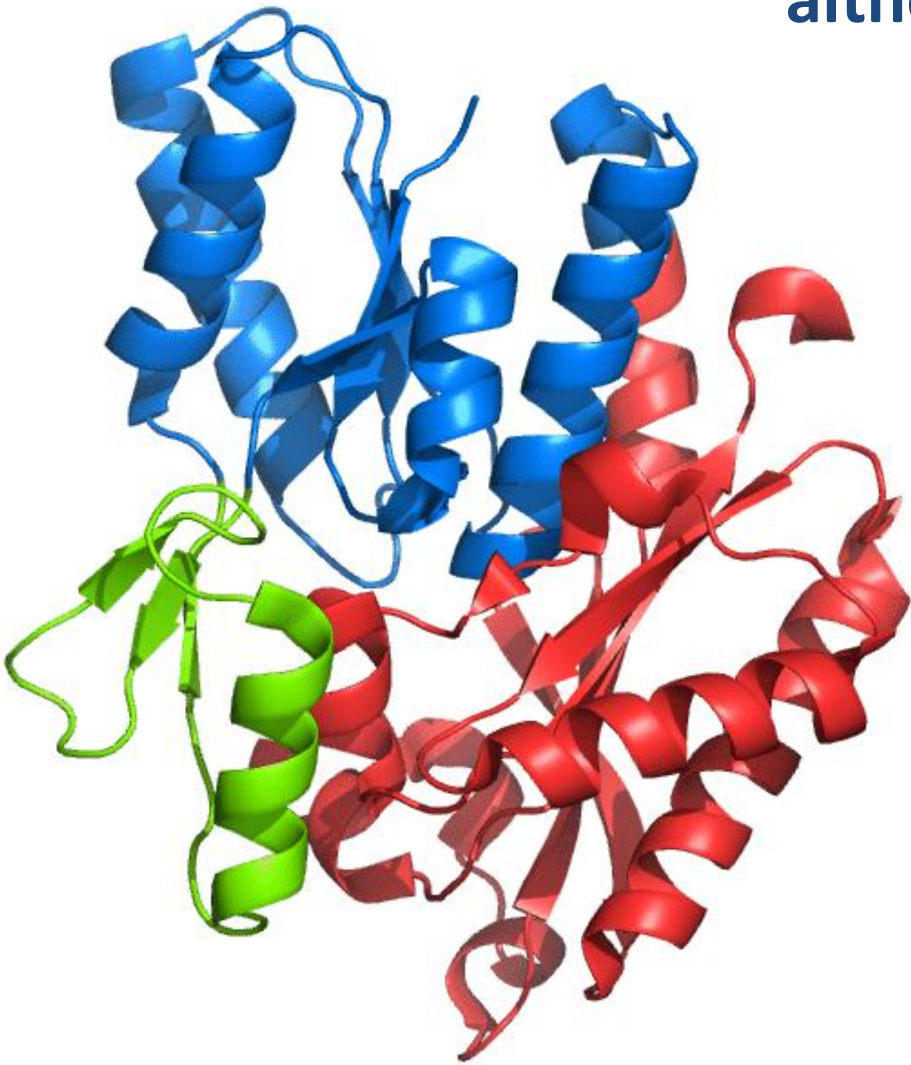
# T0535 needs domain evaluation

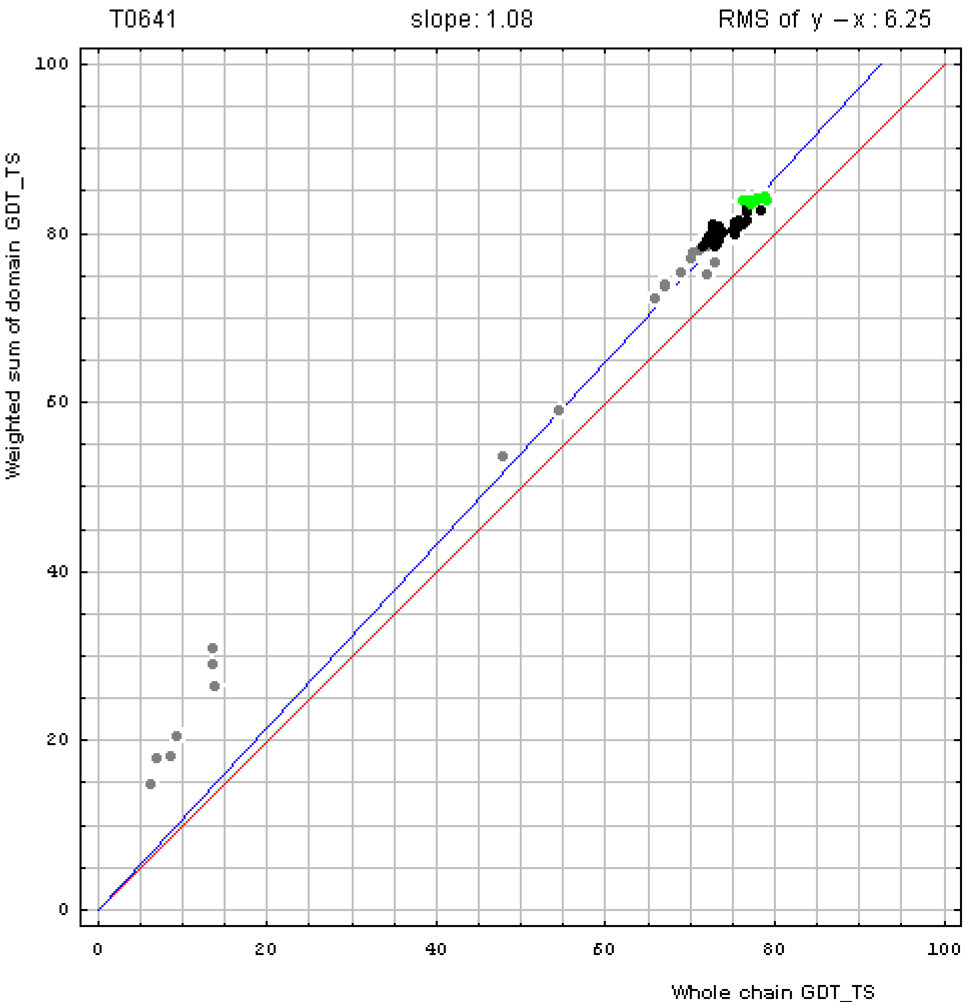




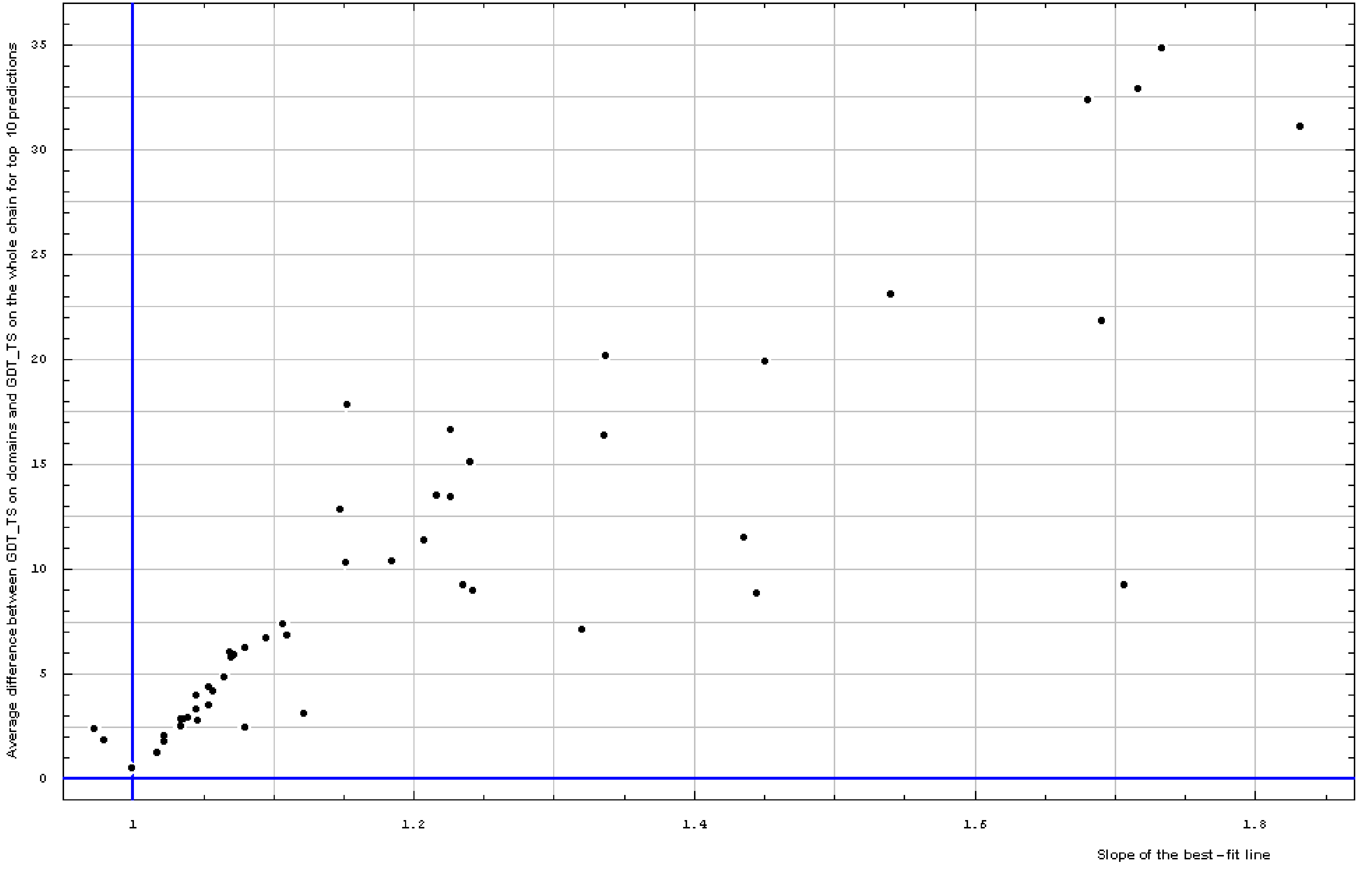

Correlation between weighted by the number of residues sum of GDT-TS scores for domain-based evaluation (y, vertical axis) and whole chain GDT-TS (x, horizontal axis).

# T641 does not need domain-based evaluation, although it consists of 3 domains

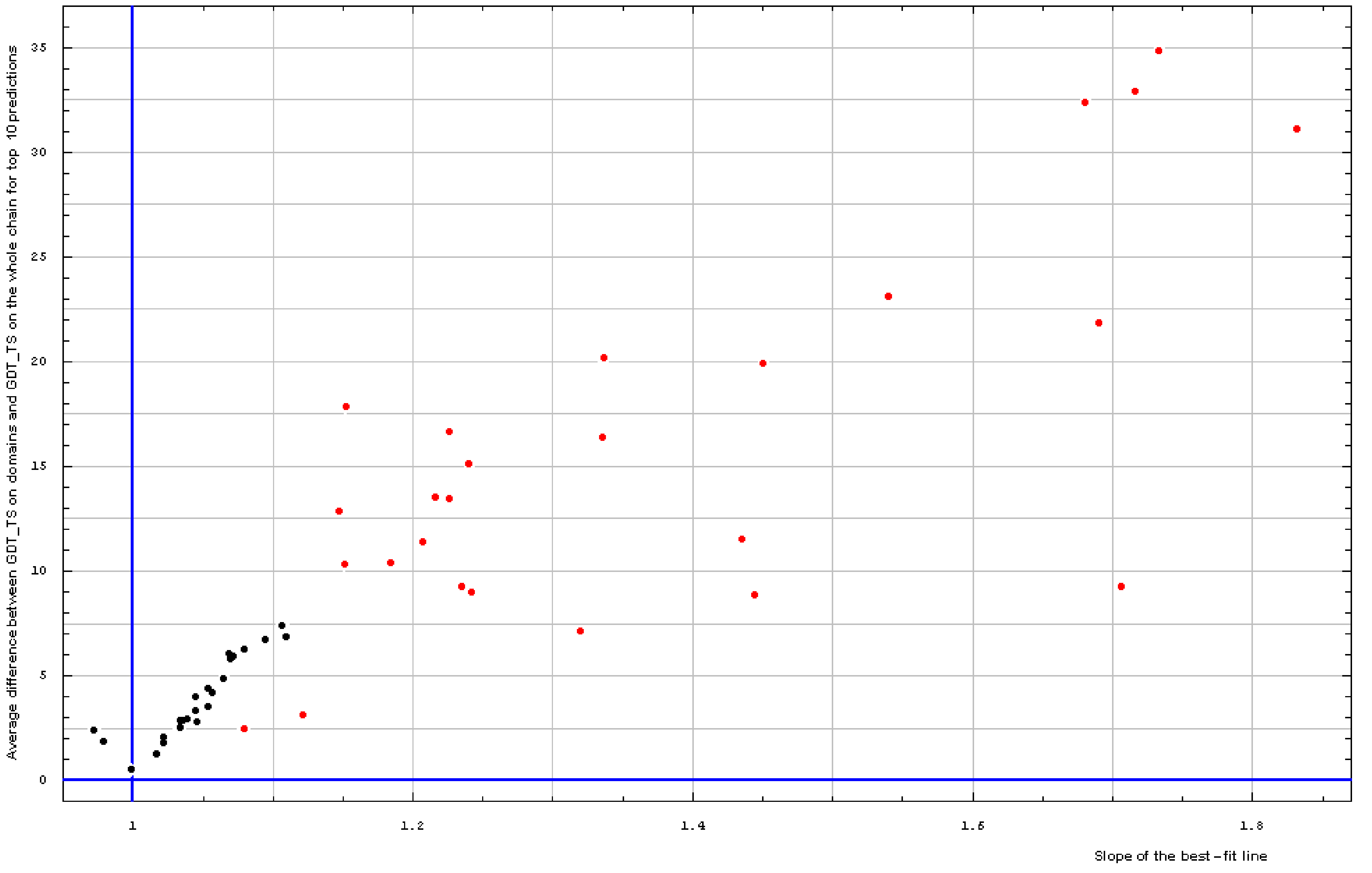




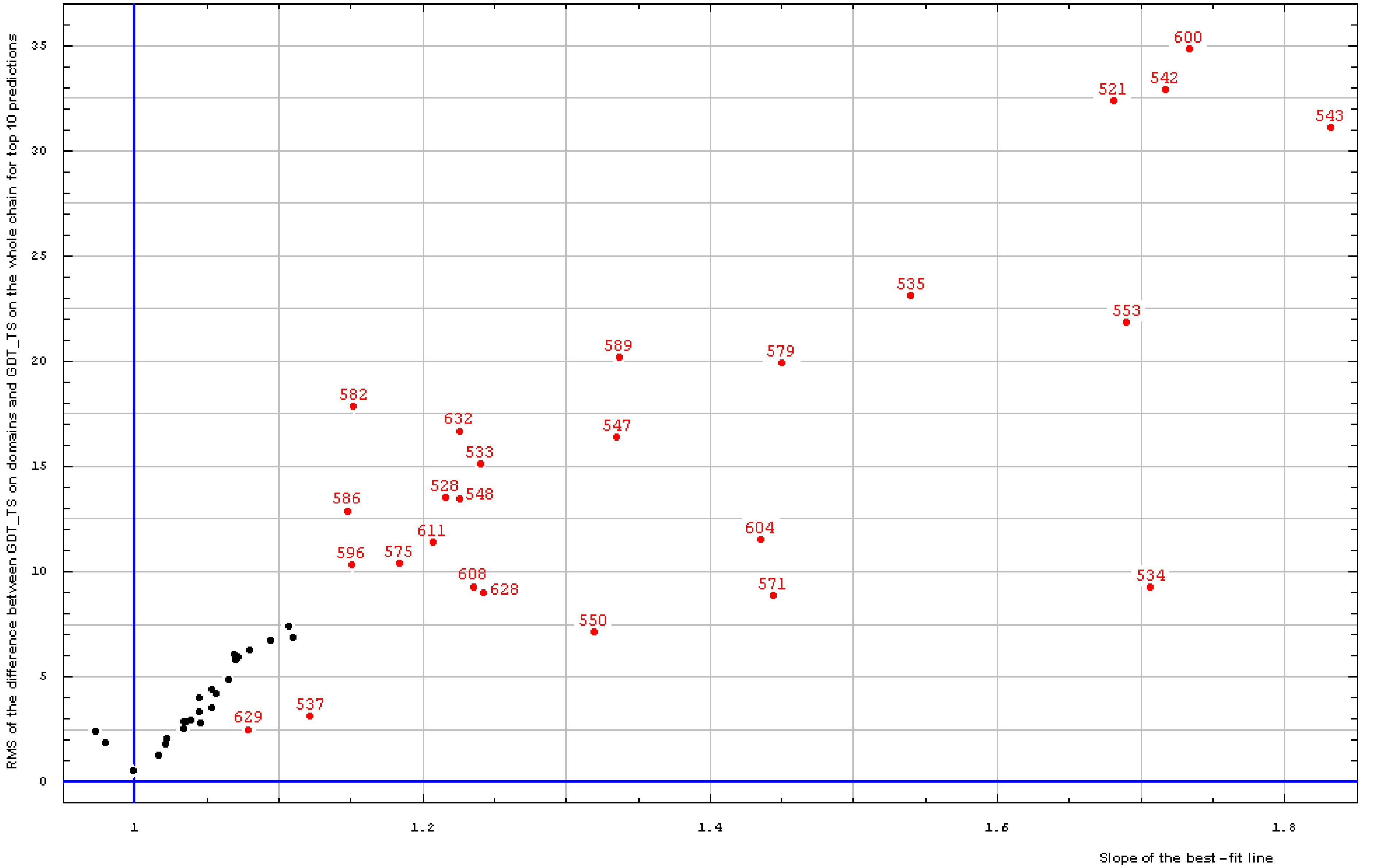

Correlation between weighted by the number of residues sum of GDT-TS scores for domain-based evaluation (y, vertical axis) and whole chain GDT-TS (x, horizontal axis).



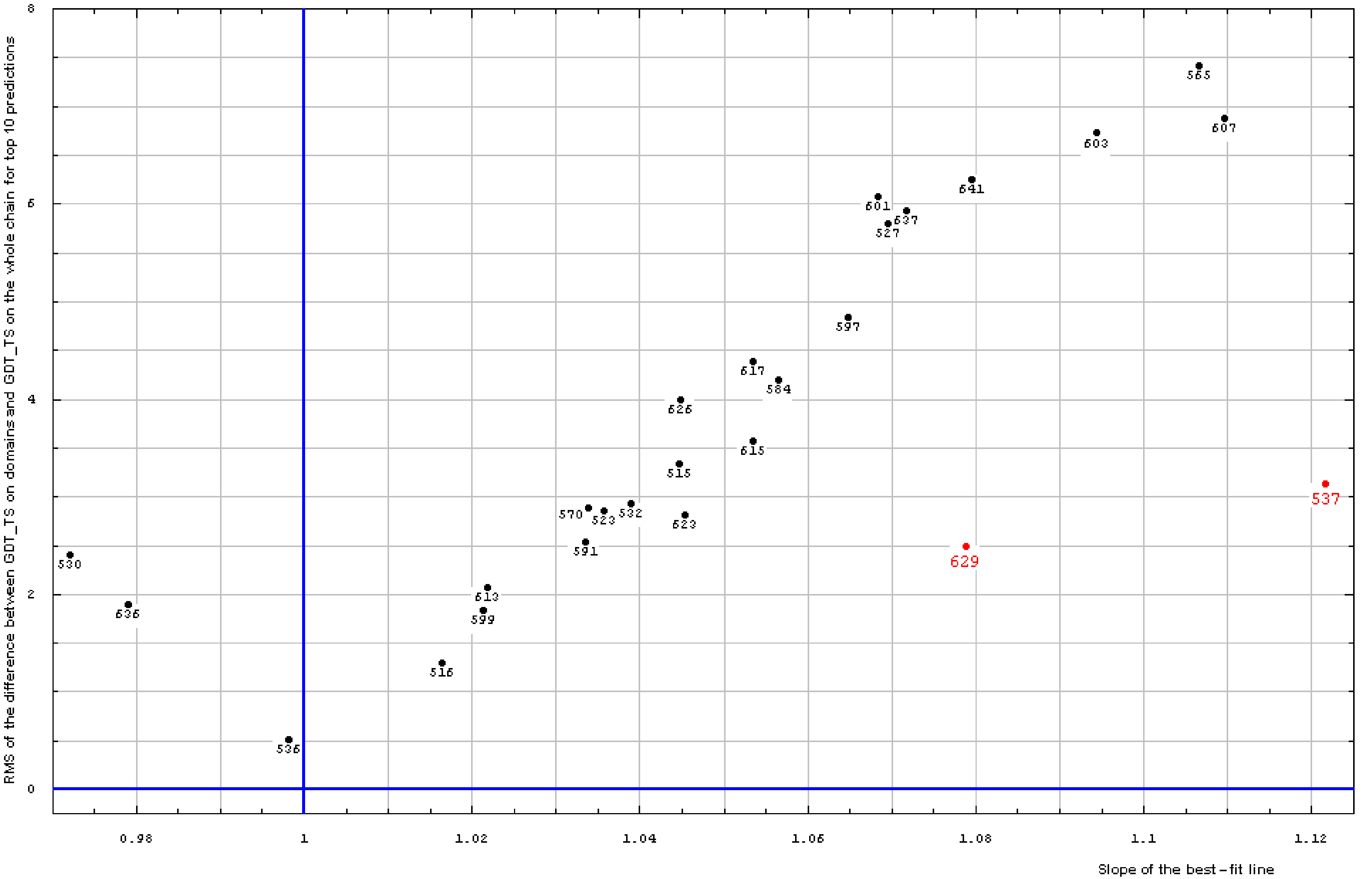

**All targets**: Correlation between RMS of the difference between GDT_TS on domains and GDT_TS on the whole chain (vertical axis) and the slope of the best-fit line (horizontal axis), both computed on top 10 server predictions.



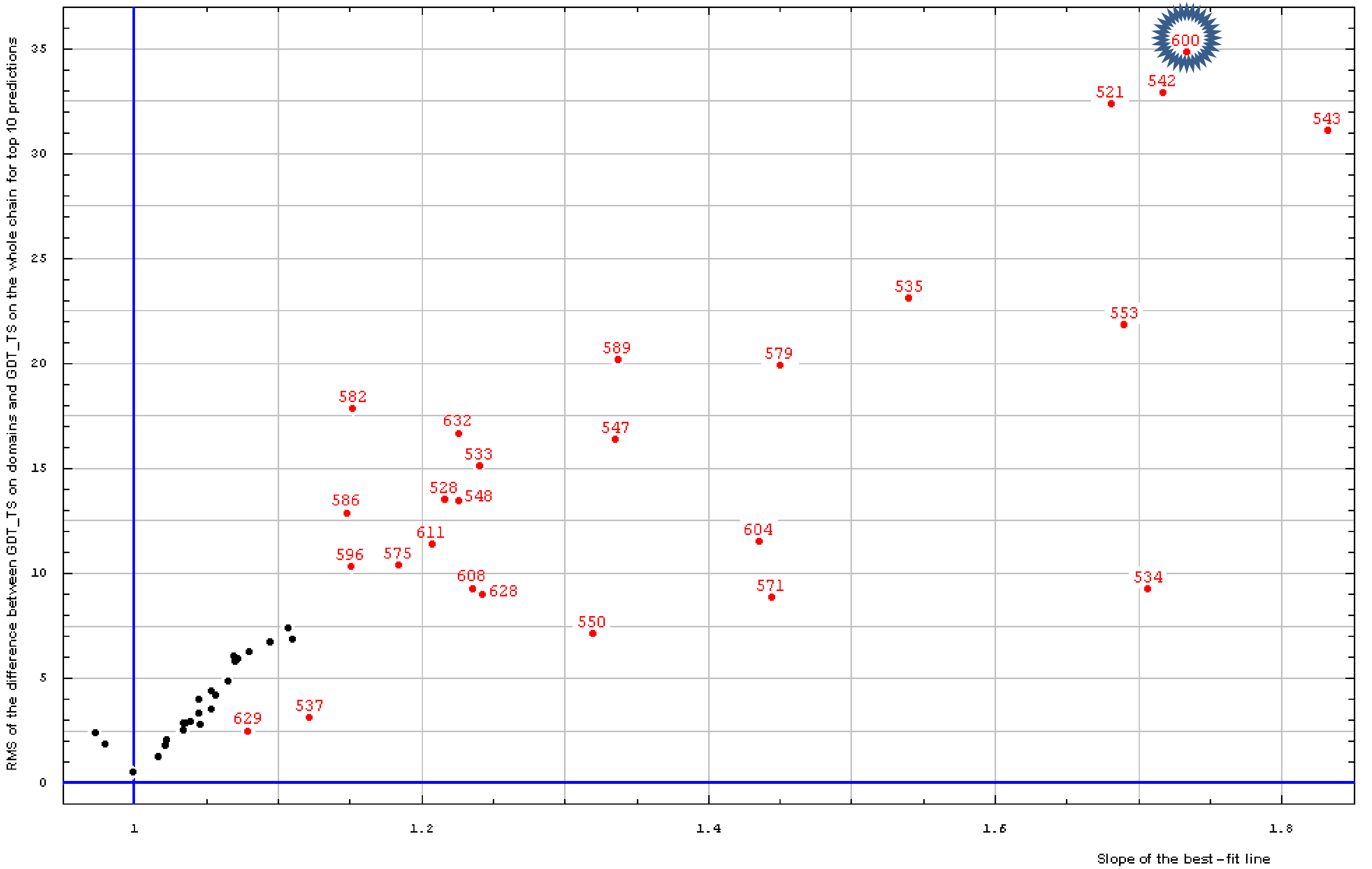

**All targets**: Correlation between RMS of the difference between GDT_TS on domains and GDT_TS on the whole chain (vertical axis) and the slope of the best-fit line (horizontal axis), both computed on top 10 server predictions.



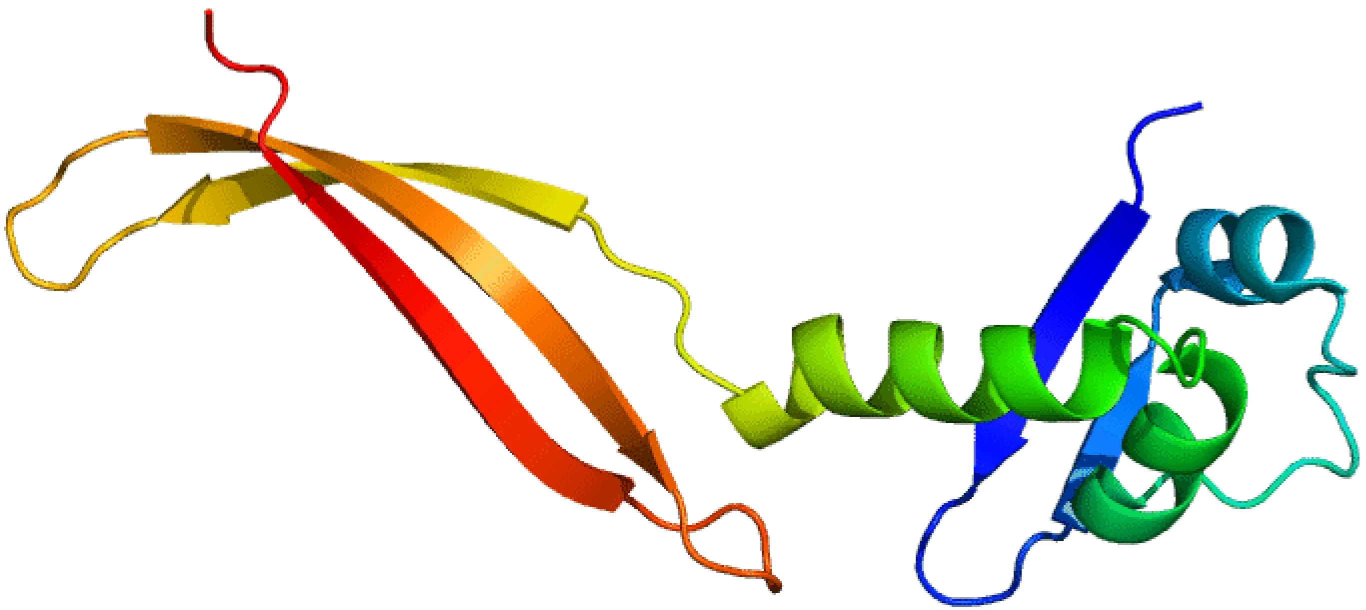

**All targets**: Correlation between RMS of the difference between GDT_TS on domains and GDT_TS on the whole chain (vertical axis) and the slope of the best-fit line (horizontal axis), both computed on top 10 server predictions.



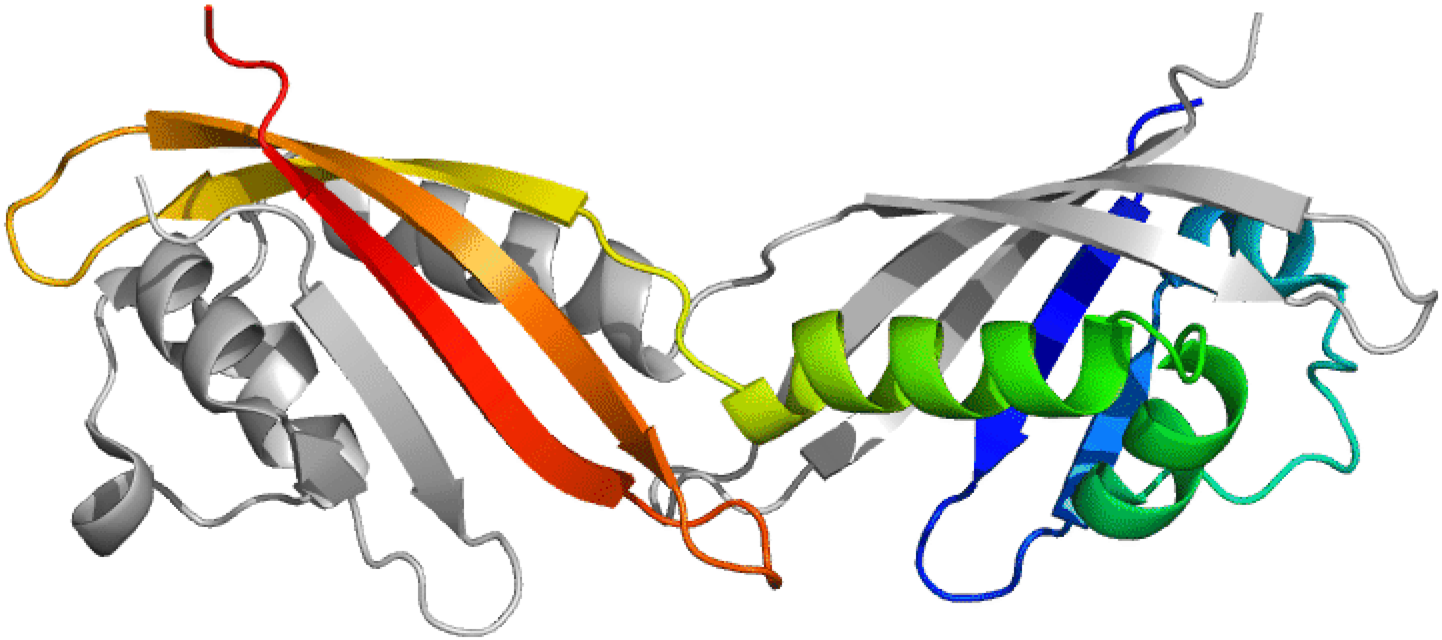

**Targets with little domain movement**: Correlation between RMS of the difference between GDT_TS on domains and GDT_TS on the whole chain (vertical axis) and the slope of the best-fit line (horizontal axis), both computed on top 10 server predictions.



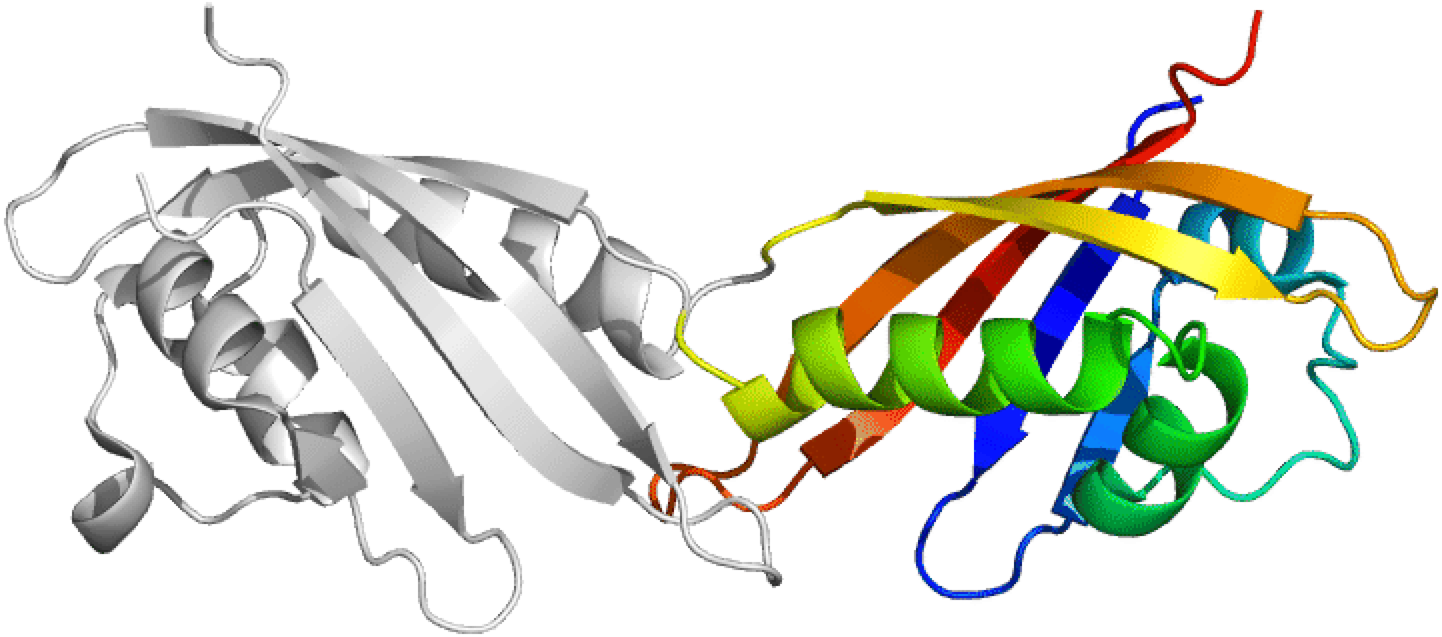

**All targets**: Correlation between RMS of the difference between GDT_TS on domains and GDT_TS on the whole chain (vertical axis) and the slope of the best-fit line (horizontal axis), both computed on top 10 server predictions.



Ribbon diagram of 600: 3nja chain A



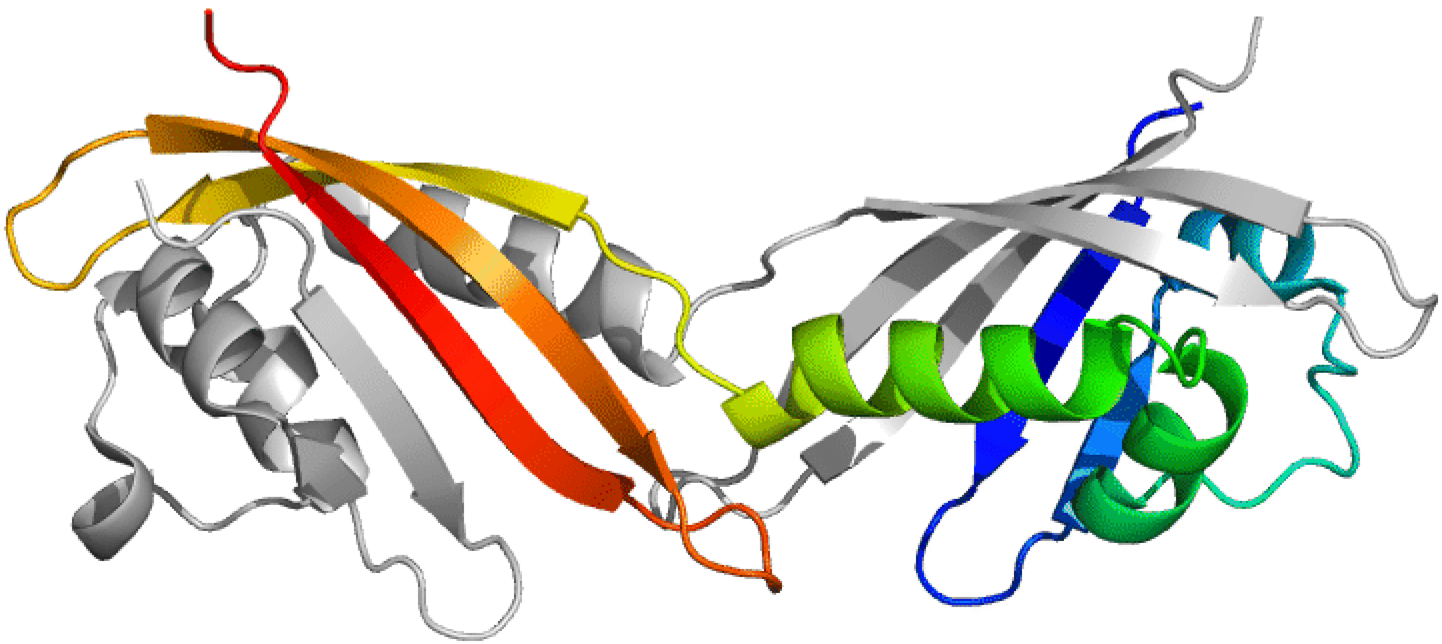
Ribbon diagram of 600: 3nja chain A



Ribbon diagram of 600: 3nja chains A and B.

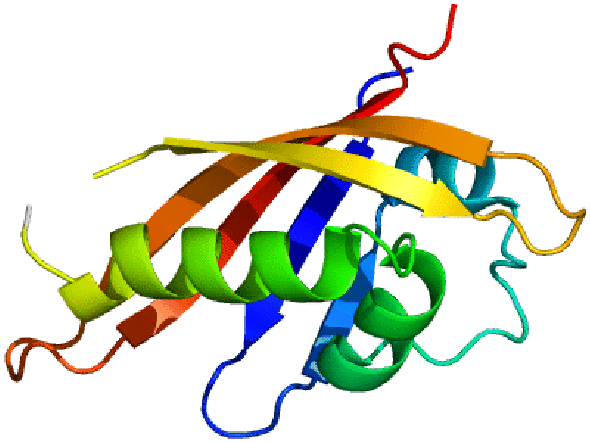
# Domain swaps!

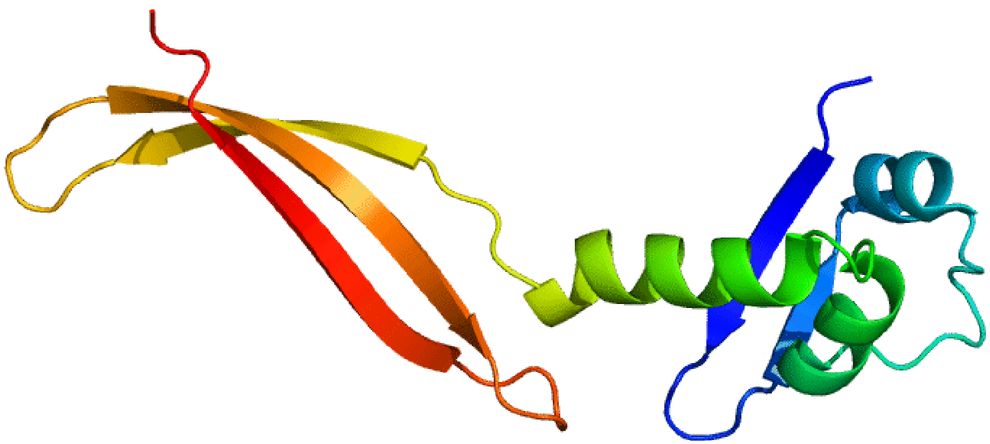
5 out of 116 targets **(4% !!!!)** exhibit domain swaps

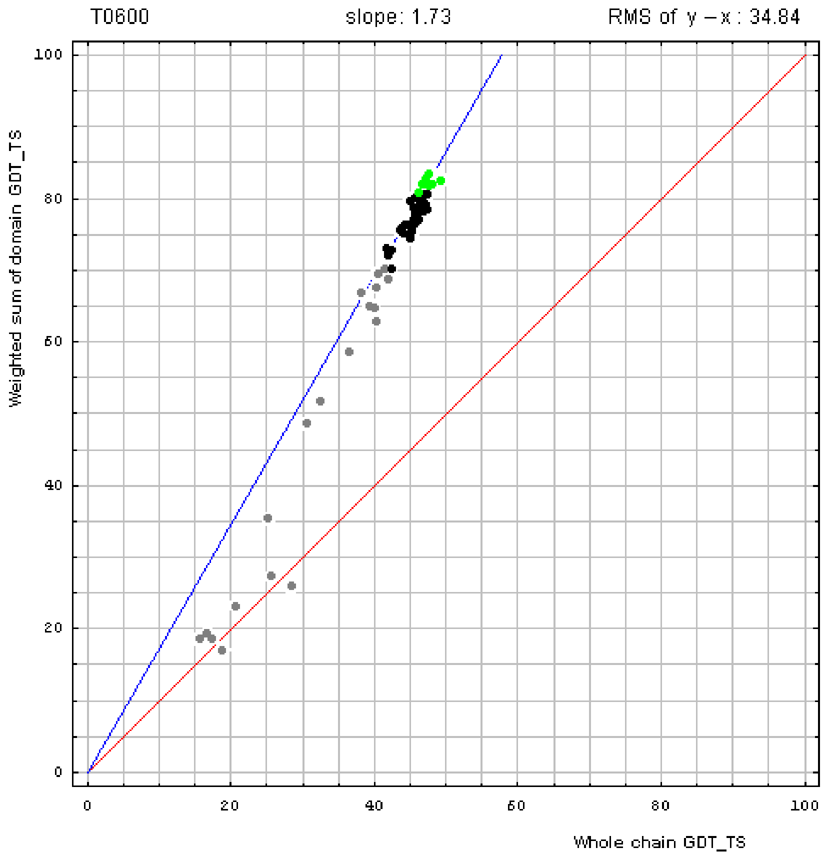


Ribbon diagram of 600: 3nja chains A and B.

# Correlation plot of swapped domain vs. full chain







# Final result:

**116** targets

**173** evolutionary domains

**146 assessment units**,

where domain split was of interest
based on the analysis of server models

# Talk plan

- Target Overview
- Domain Definition
- Domain Classification
- CASP9 categories: TBM and FM

# Target Classification

**1. biology**, i.e. evolutionary classification

**2. assessment**, i.e. CASP category classification

# Evolutionary Classification of targets

We **find** if any proteins with known structures are **homologous** to CASP targets, their domains and domain combinations

How is it relevant to structure prediction and CASP? you might ask, my dear friend.

And the answer is:
it is as relevant as any **biological information**

you might think that you don't need it,
but then you would start wondering
why your predictions look like crap …

# Evolutionary Classification of targets

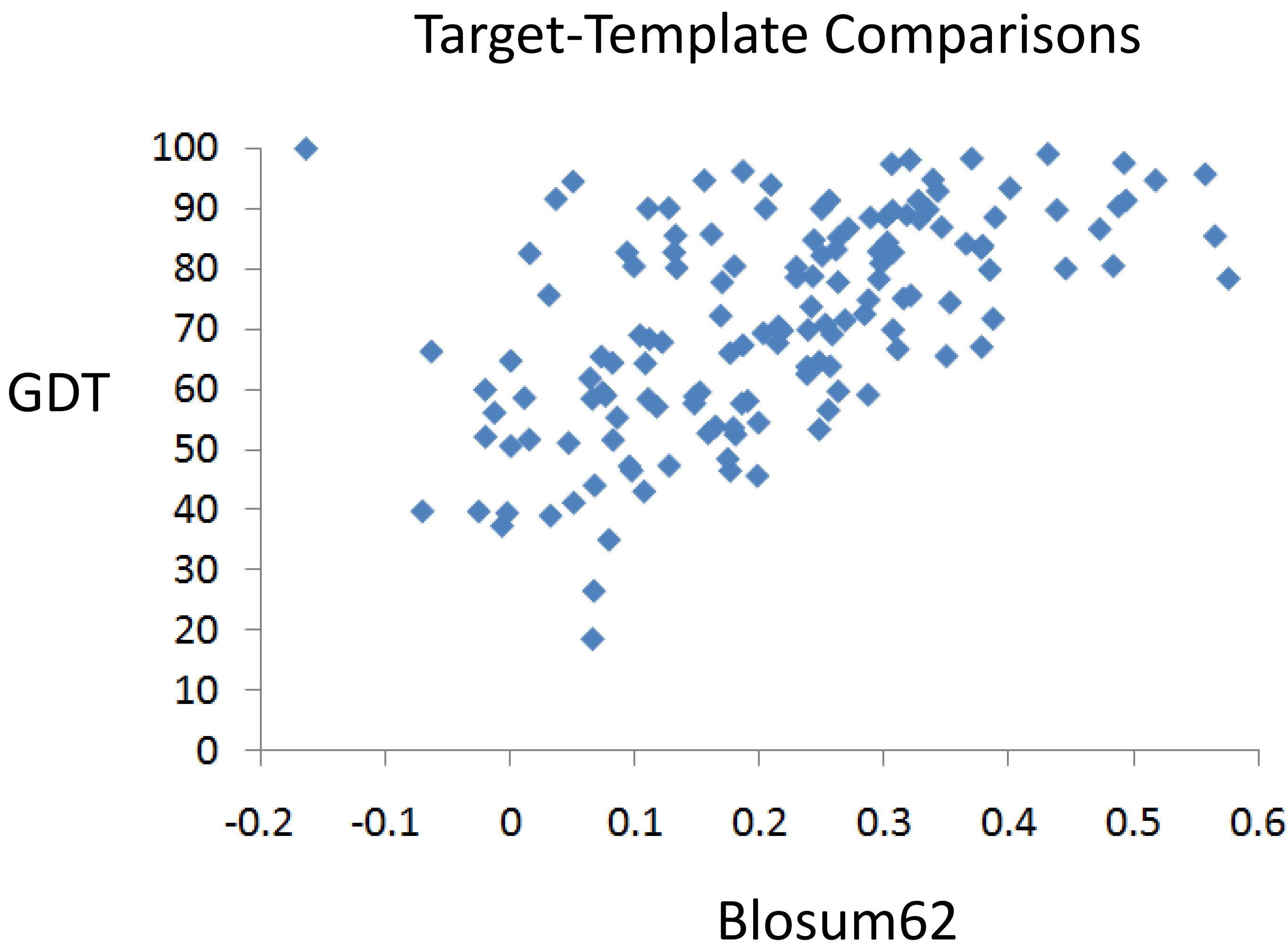
The best indication of homology is statistically significant and meaningful sequence/profile similarity found **prior** to knowledge of 3D structure:
i.e. predictions are relevant for evolutionary classification

**1.** During CASP season, we had "spies" in the group, who were running predictions to see what can be done without structural knowledge (PSI-BLAST, HHsearch)

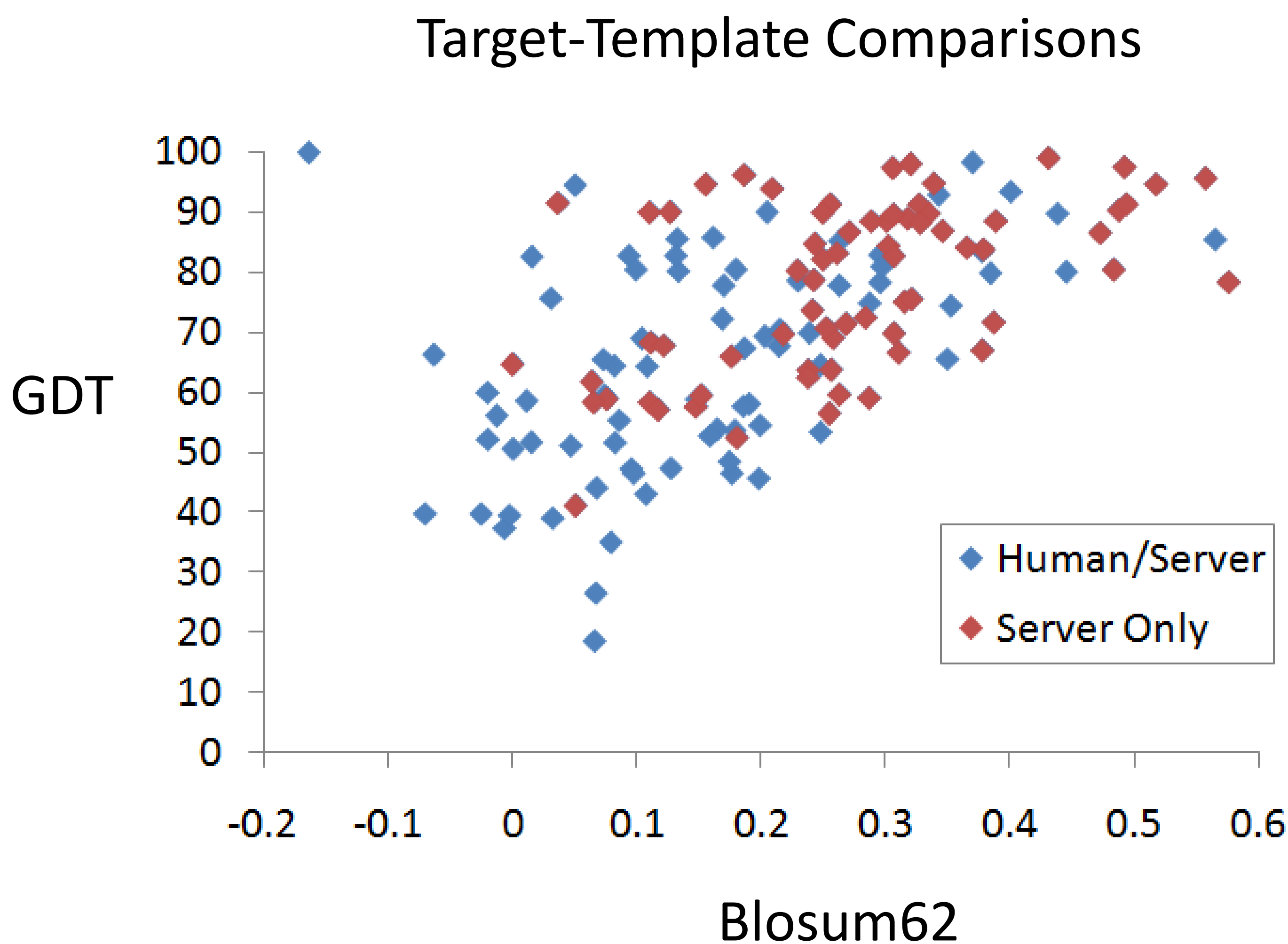
**2.** After 3D structures became available, we searched PDB for matches to target structures (DALI, TM-align, LGA)

**3.** Analyzing the results of **1** and **2** we found quite a few interesting things about CASP targets

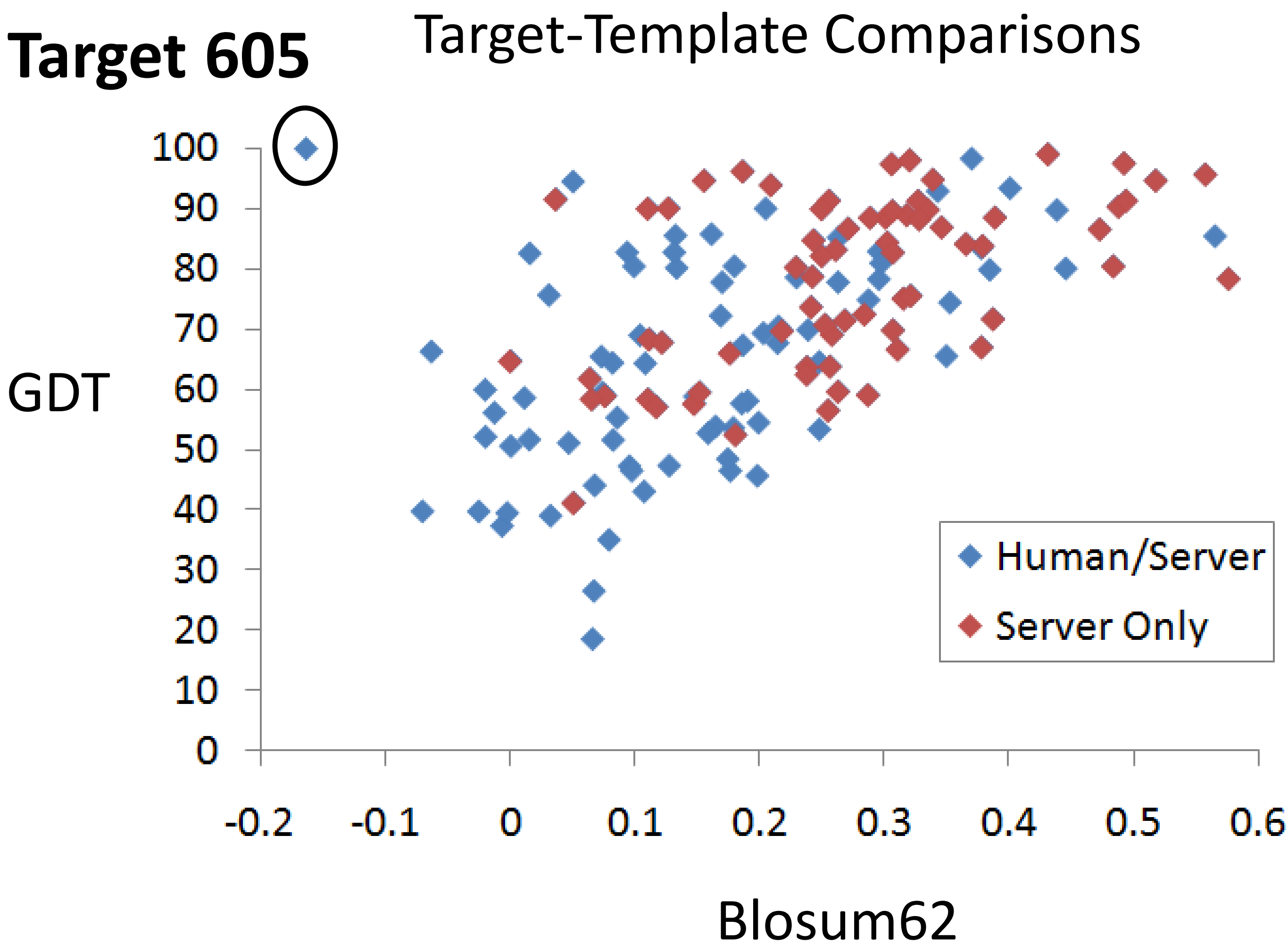
# CASP9 Target Distribution



# CASP9 Target Distribution

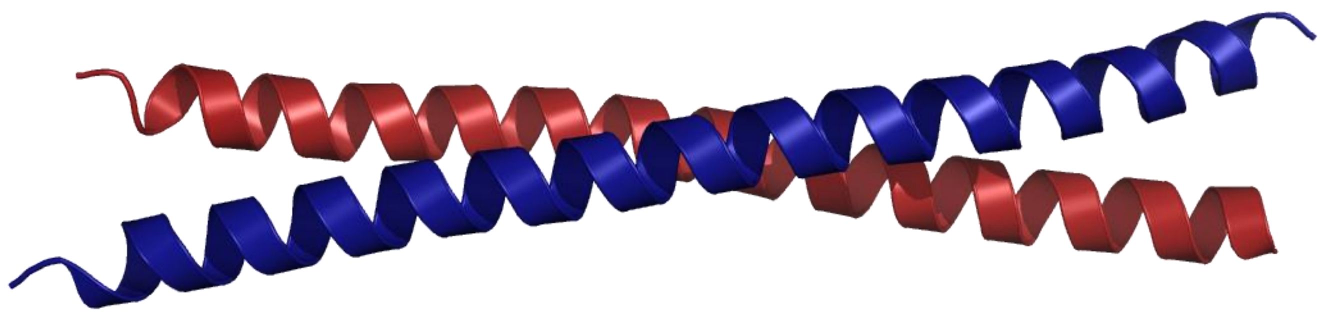


# CASP9 Target Distribution



# CASP9 Target Distribution

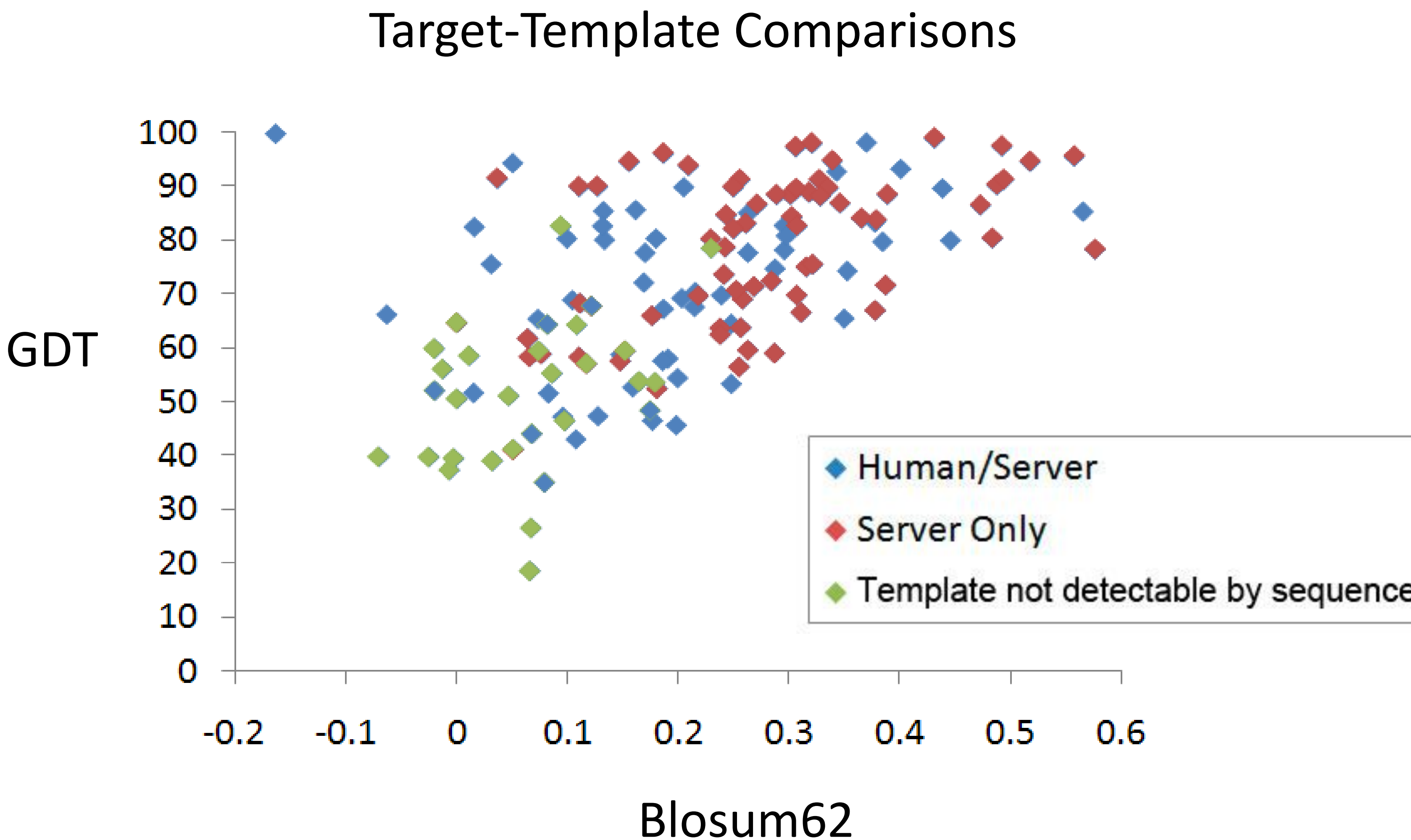
**Target 605**



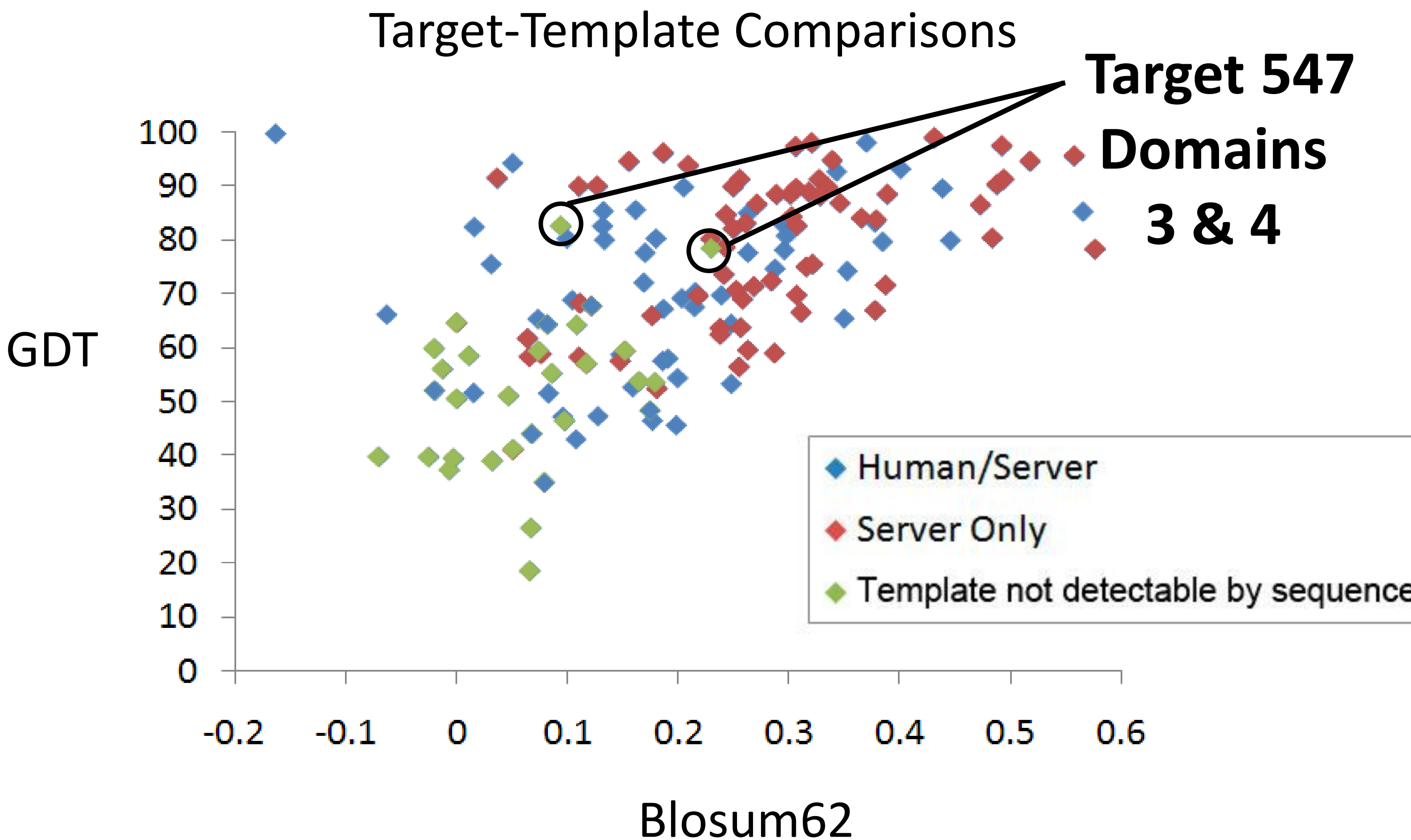
**>T0605 3NMD, unknown species, 72 residues**

MRGSHHHHHHGMASIEGRGSLRDLQYALQEKIEELRQRDALIDELELELDQKDELIQMLQNELDKYRSVIRP

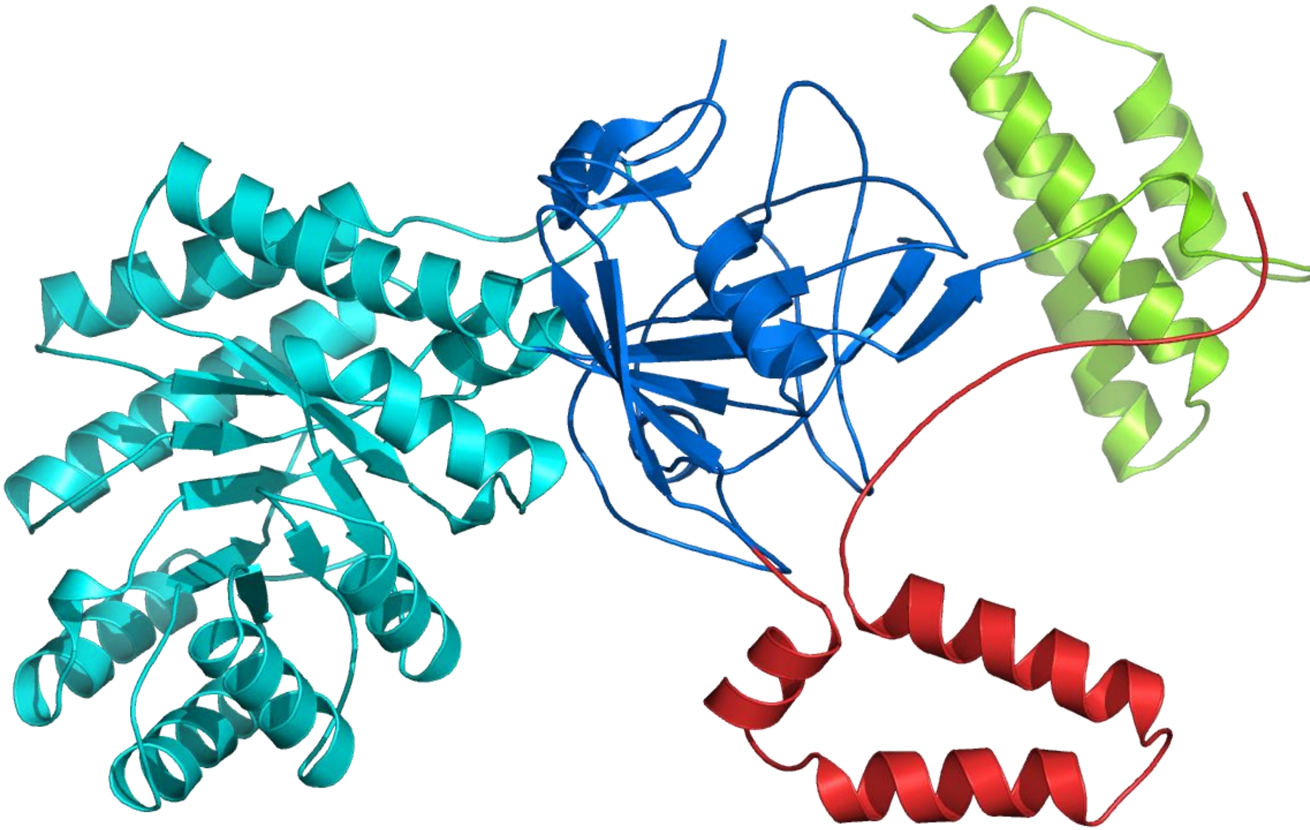
# CASP9 Target Distribution



# CASP9 Target Distribution

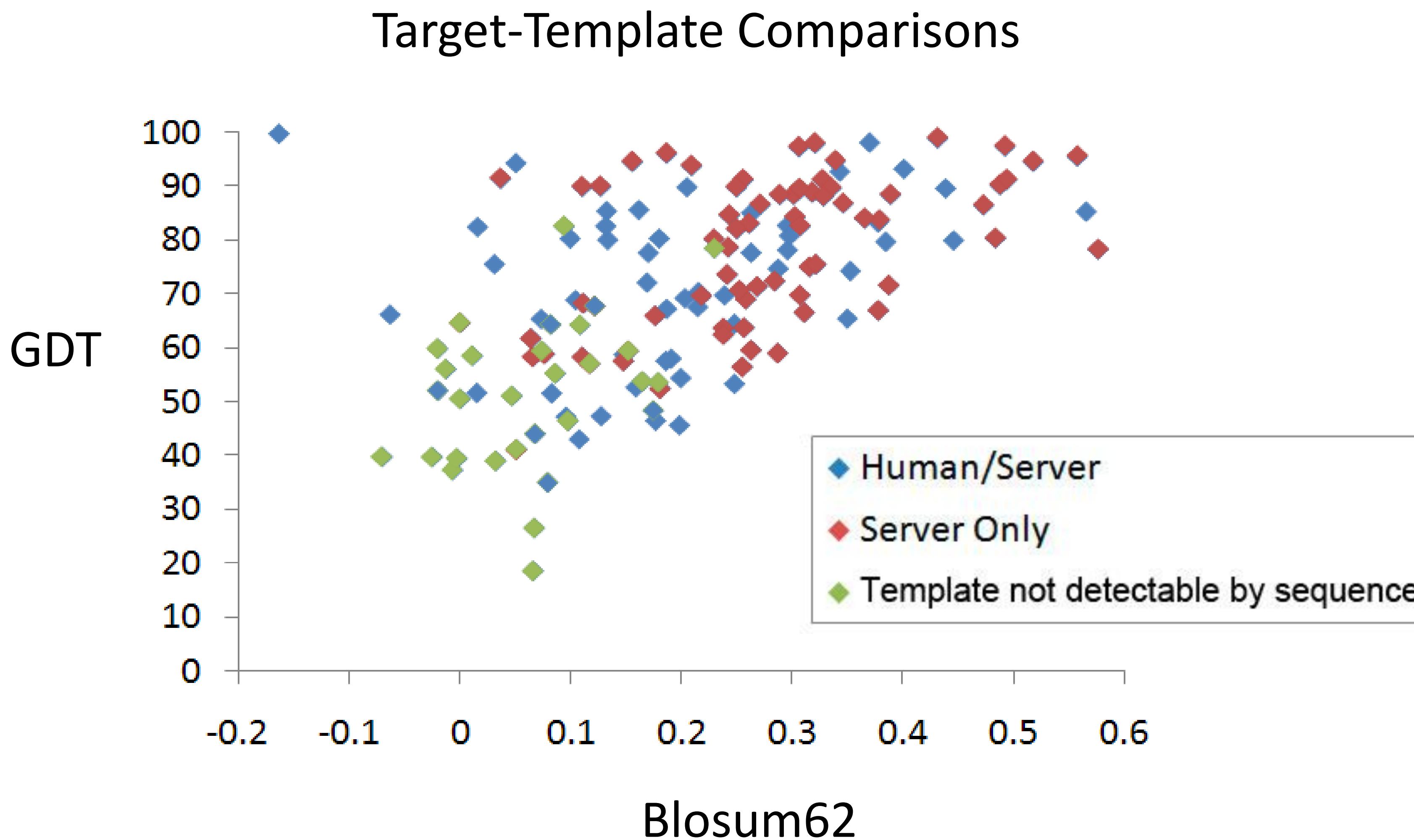


# Target 547

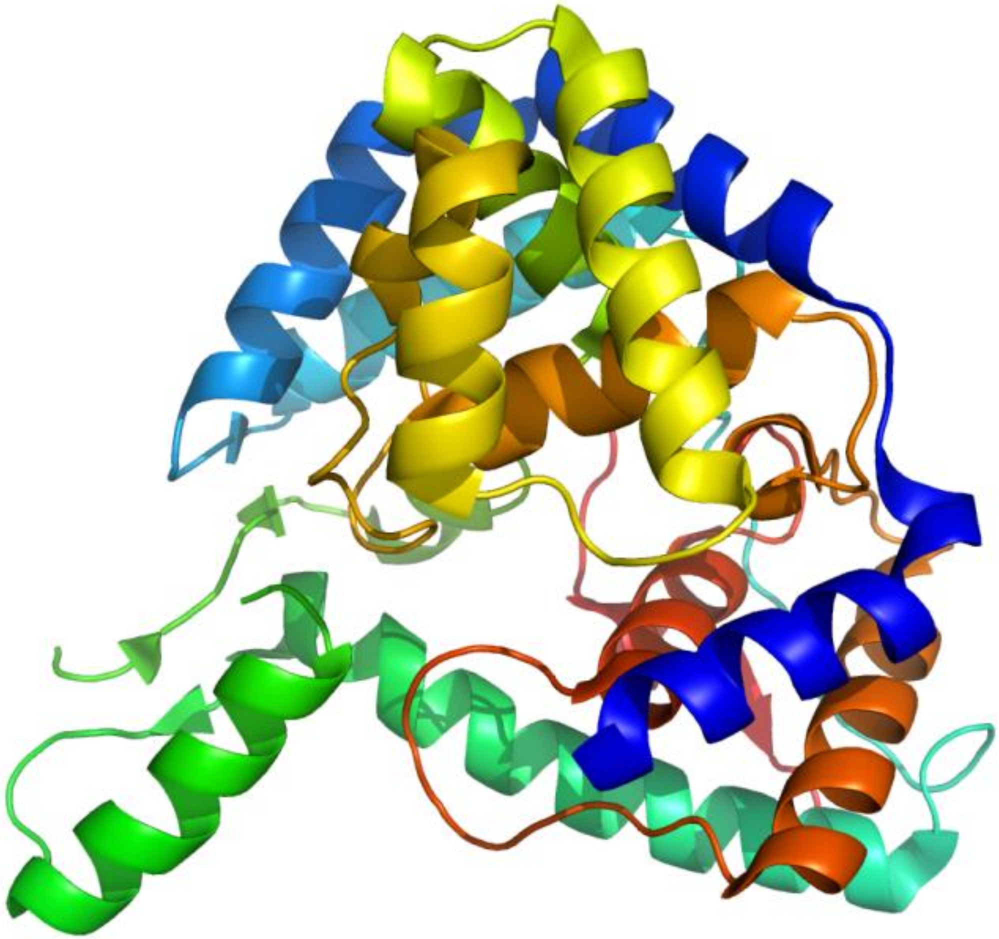


**Target 547 Domains 3 & 4**

# CASP9 Target Distribution

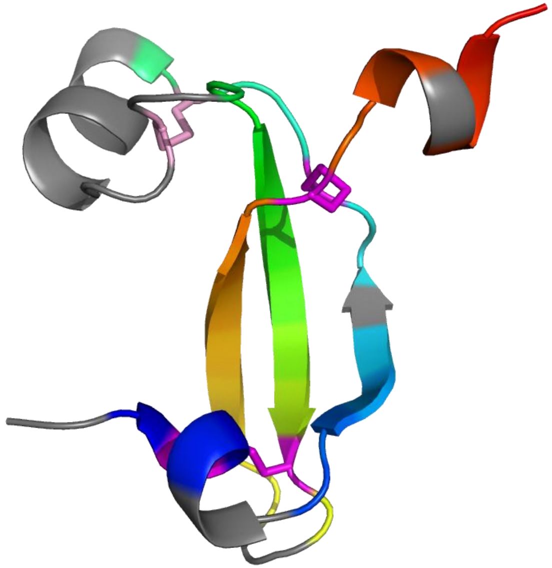


# 529d1: a new fold?

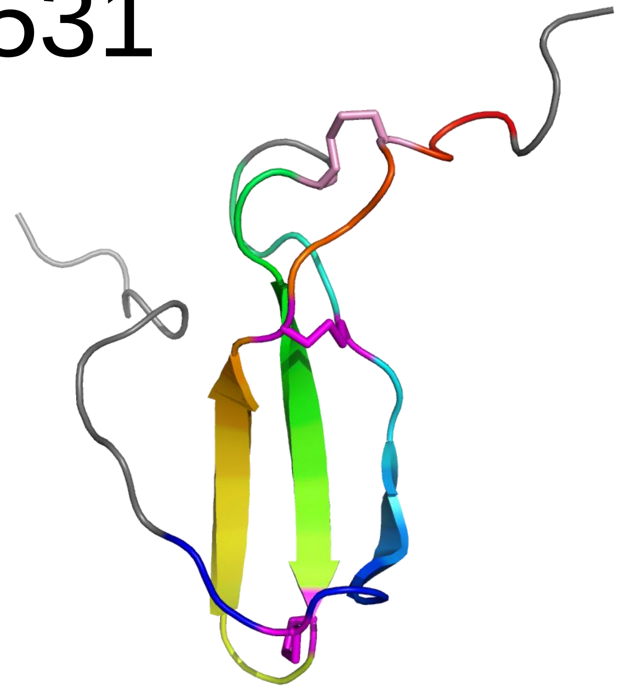


No similar structures found. Highest Dali Z score 2.1.

# Target 531



Dali Z
1.6



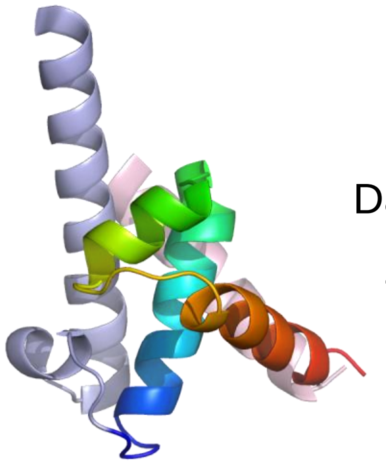
531: Jumping translocation breakpoint protein, extracellular domain

Midkine: a heparin-binding growth factor, N-terminal domain (1mkn)

```
 531   gsgmkefPCWLveEFVVaEECSPCSnfrakttpecgpTGYVEKITCSssKRNEFKSCRSAlMEQR
1mkn   vkkggpgSECA--EWAW-GPCTPSS---------kdcGVGFREGTCG--AQTQRIRCRVP-CNWK
```

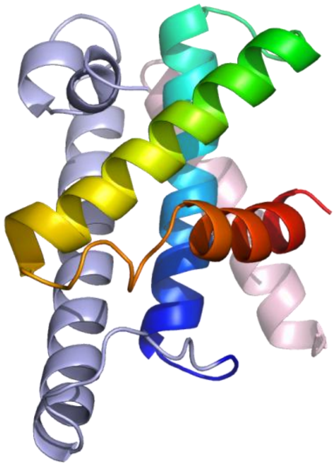
Low Dali Z (1.6), but preserves two of the three disulfides pairs.

# Target 561: an elaborated HTH?



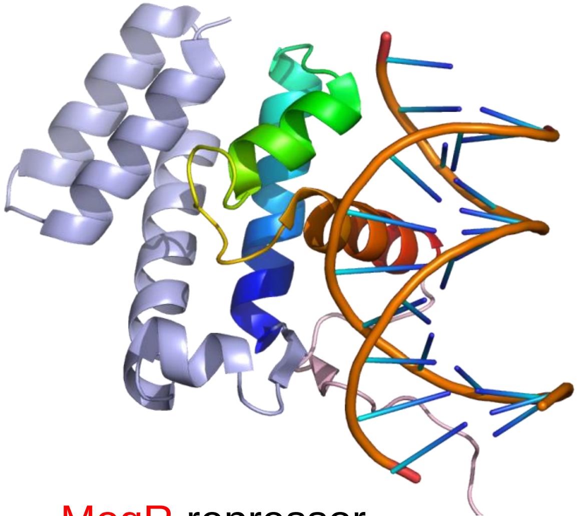
Dali Z

4.1



Dali Z

5.4



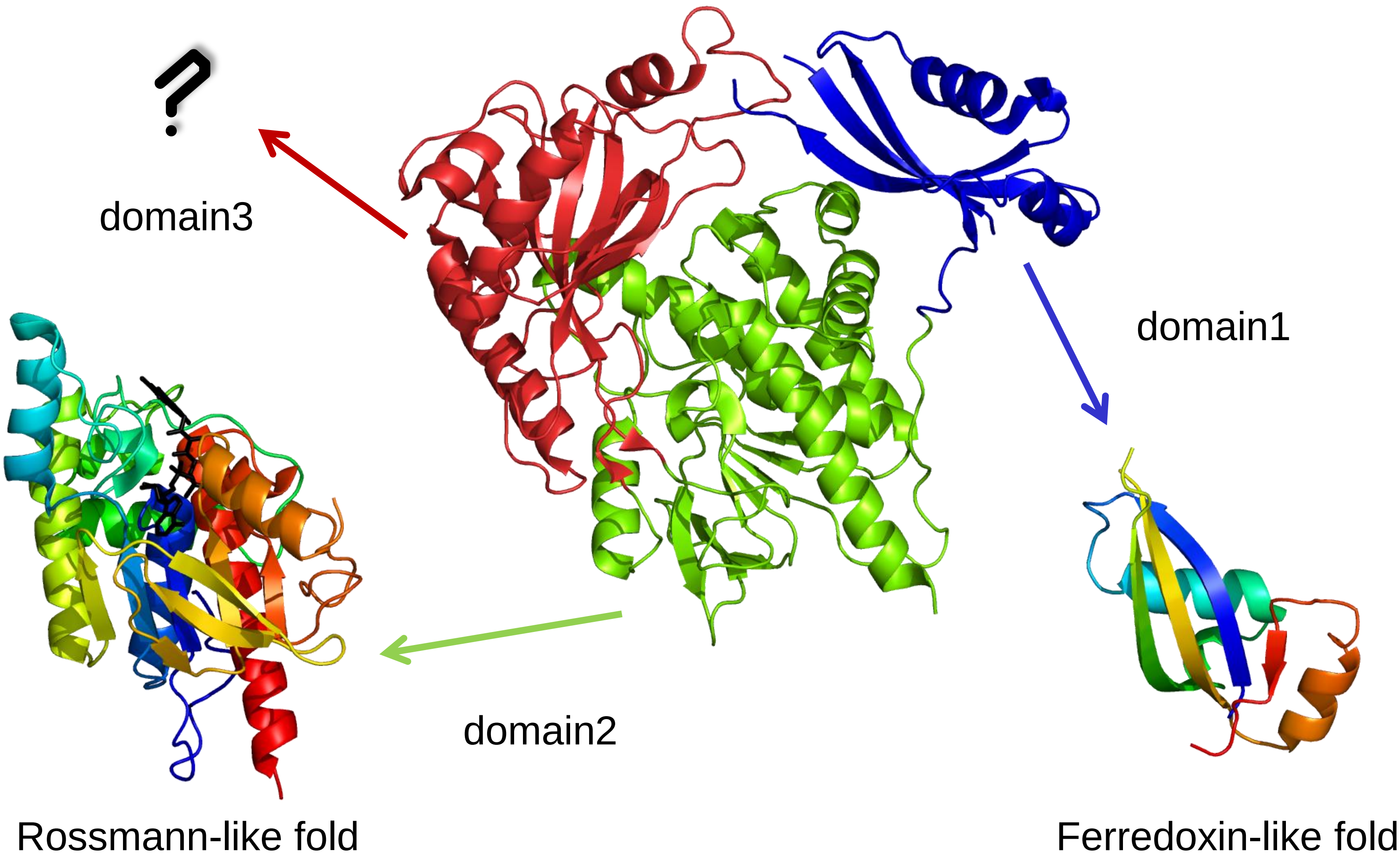
Replication initiation factor DnaA ,C-terminal domain (d1l8qa1)

561, DnaJ binding protein

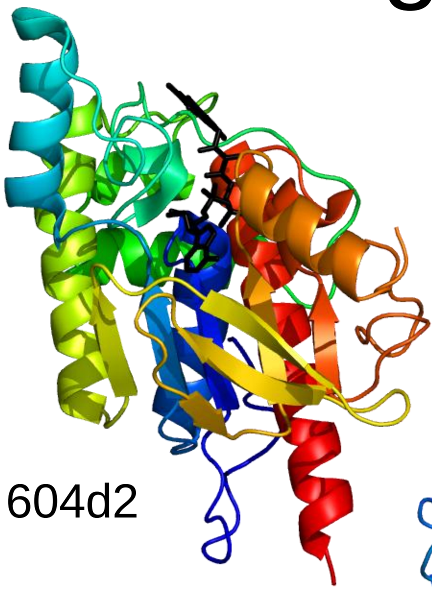
MogR repressor (3fdq)

```
 561     1 avekkkyldsEALLhcisaqlldmWKQARA-------rwLELVgkewahmlalnperkdf
1l8q     1 ---gfeglerKERK------erdkLMQIVEfvanyyavkVEDI----------------l

 561    54 lWKNQSEMNSAFFDLCEVG-KQVMlgllgkevalpkeeqaFWIMYAVHLSAacaeelhmp
1l8q    36 sDKRNKRTSEARKIAMYLCrKVCS---------------aSLIEIARAFKR---------

 561   113 evamSLRKLNVKLKDFNF-mpPEEKKRRMERKQRIEEARRhgmp 155
1l8q    72 ---kDHTTVIHAIRSVEEekkRKFKHLVGFLEKQAFDKIC---- 108
```

# Target 604: Domain Organization

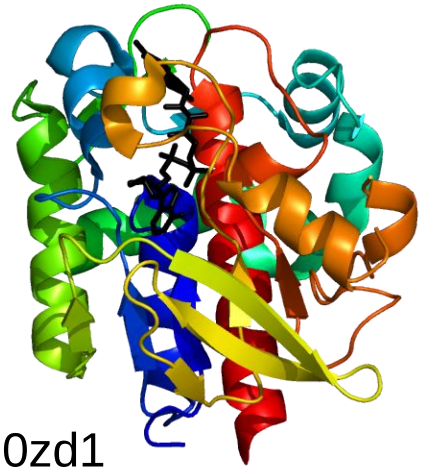


# Target 604d3: a surprise



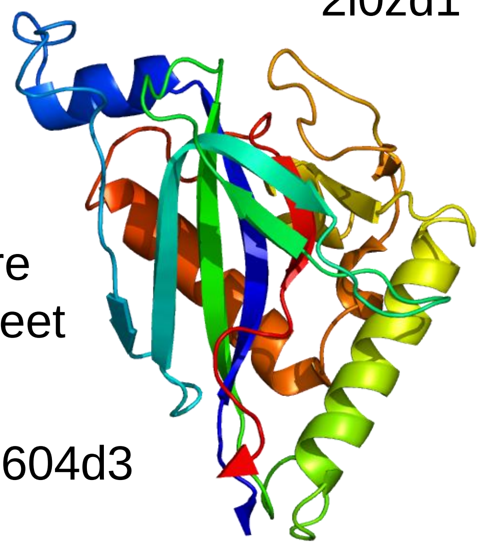
604d2

Dali Z
———
21



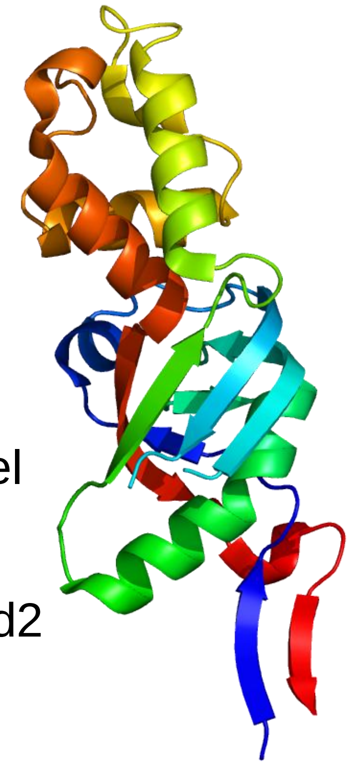
2i0zd1

Core
β-sheet



604d3

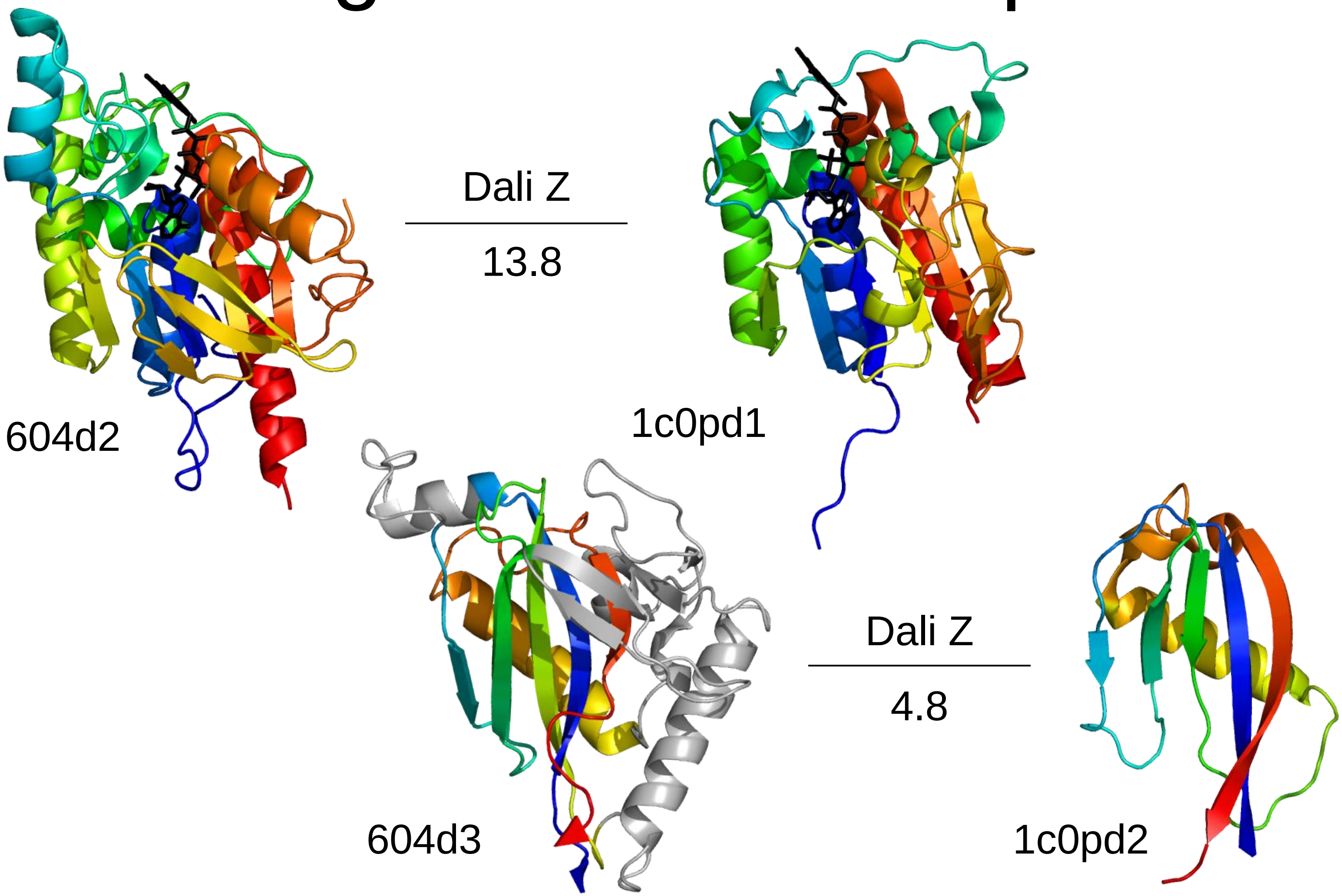
Core
β-barrel



2i0zd2

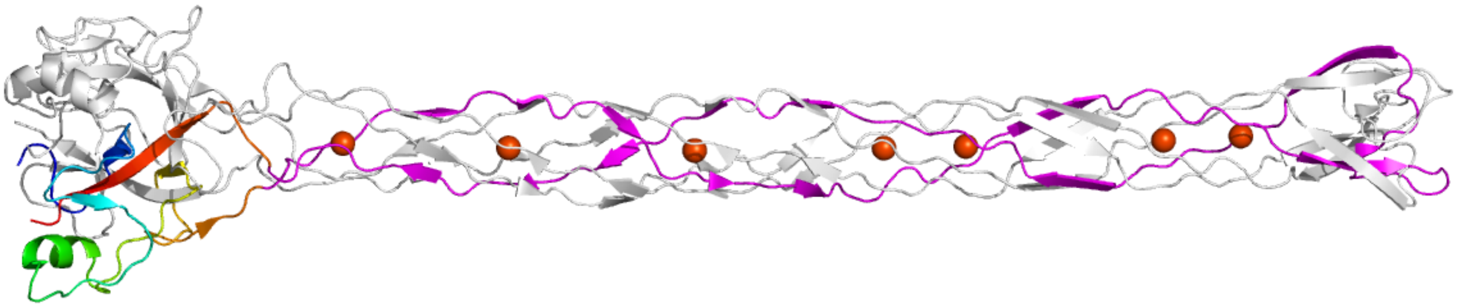
2i0z is a deceiving template: HHsearch probability is 100, and the alignment covers both domains

# Target 604d3: a surprise

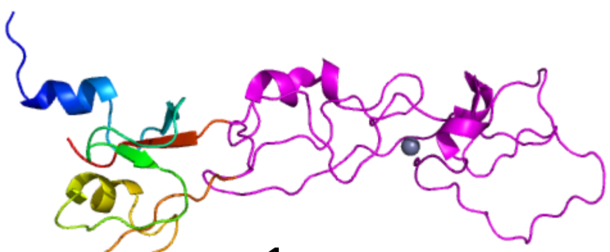



1c0pd2 is a better template for 604d3, which includes many difficult insertions.

# Target 629d2: an unusual fold



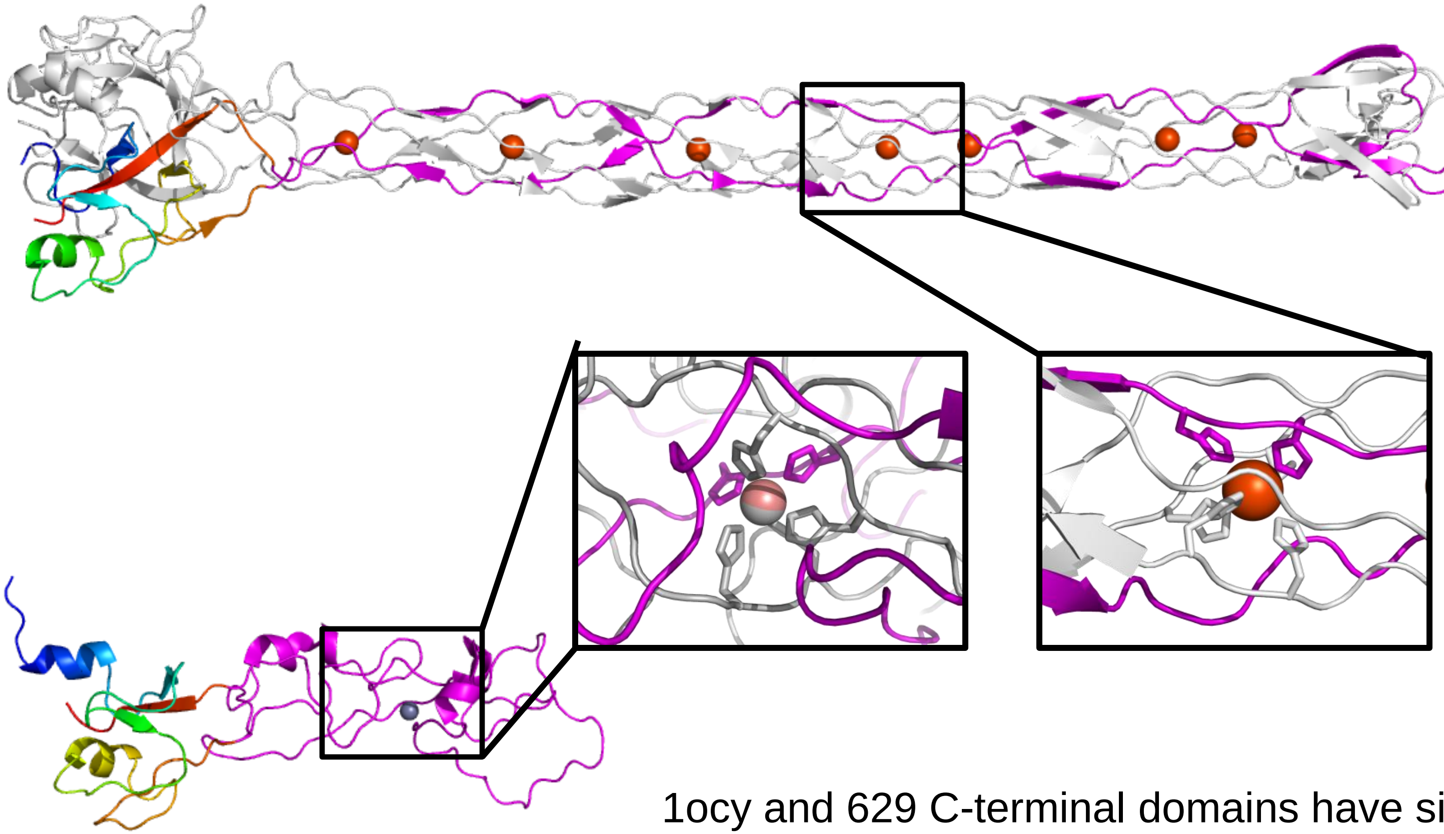
629 trimer



1ocy

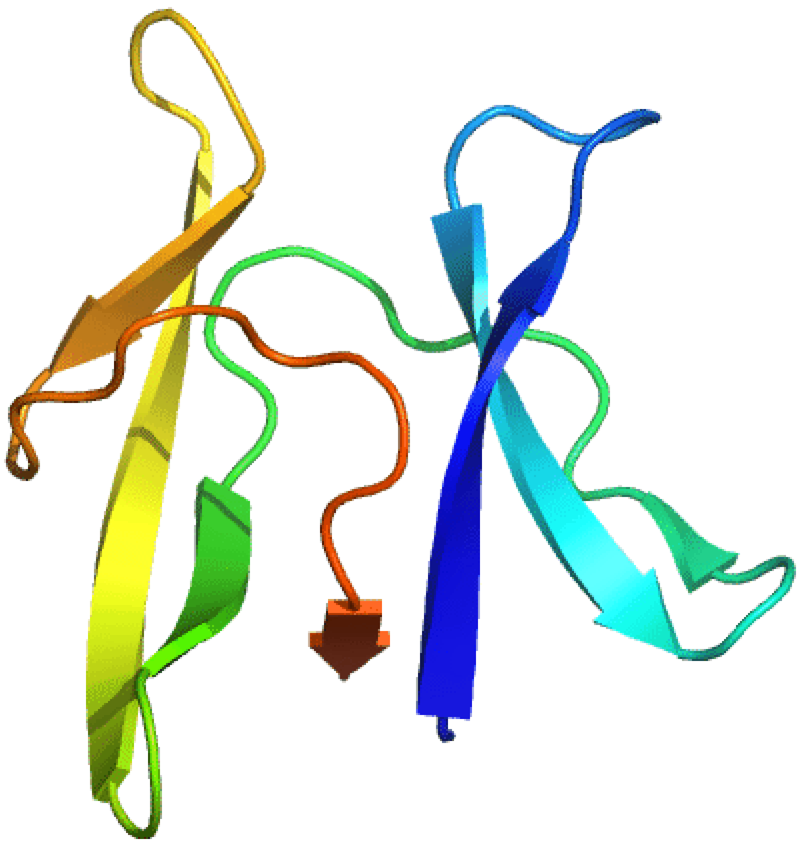
629 domain 1 is similar to 1ocy N-terminal domain, but C-terminal domains are very different

# Target 629d2: an unusual fold

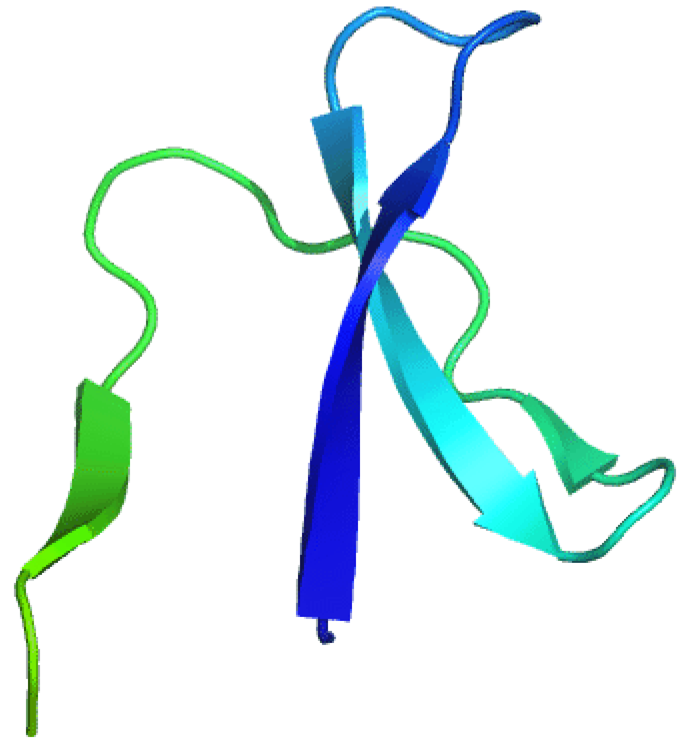


1ocy and 629 C-terminal domains have similar metal-binding sites comprised of three HXH motifs, one from each monomer. 1ocy has one metal-binding site while T0629 has seven.

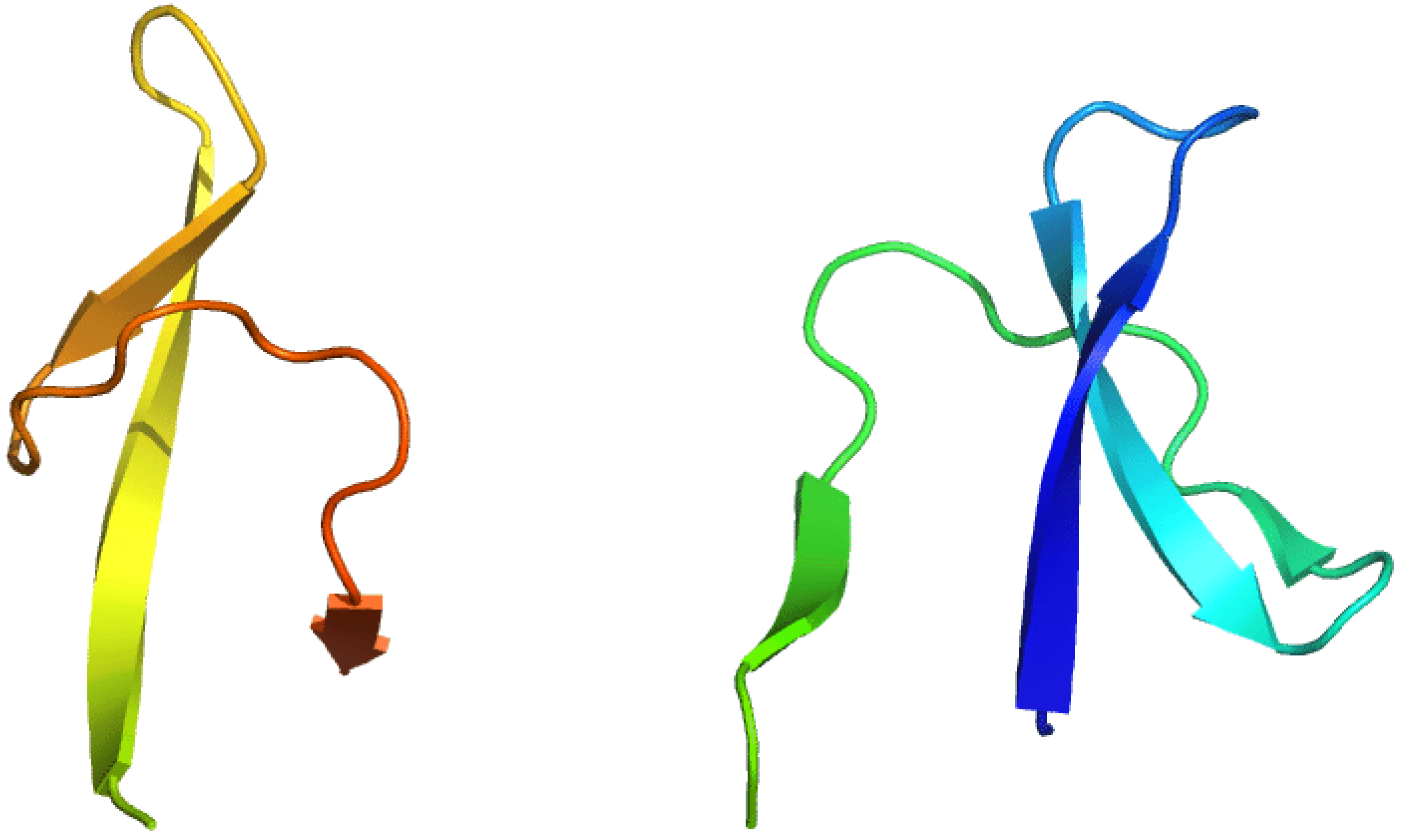
# Target 624



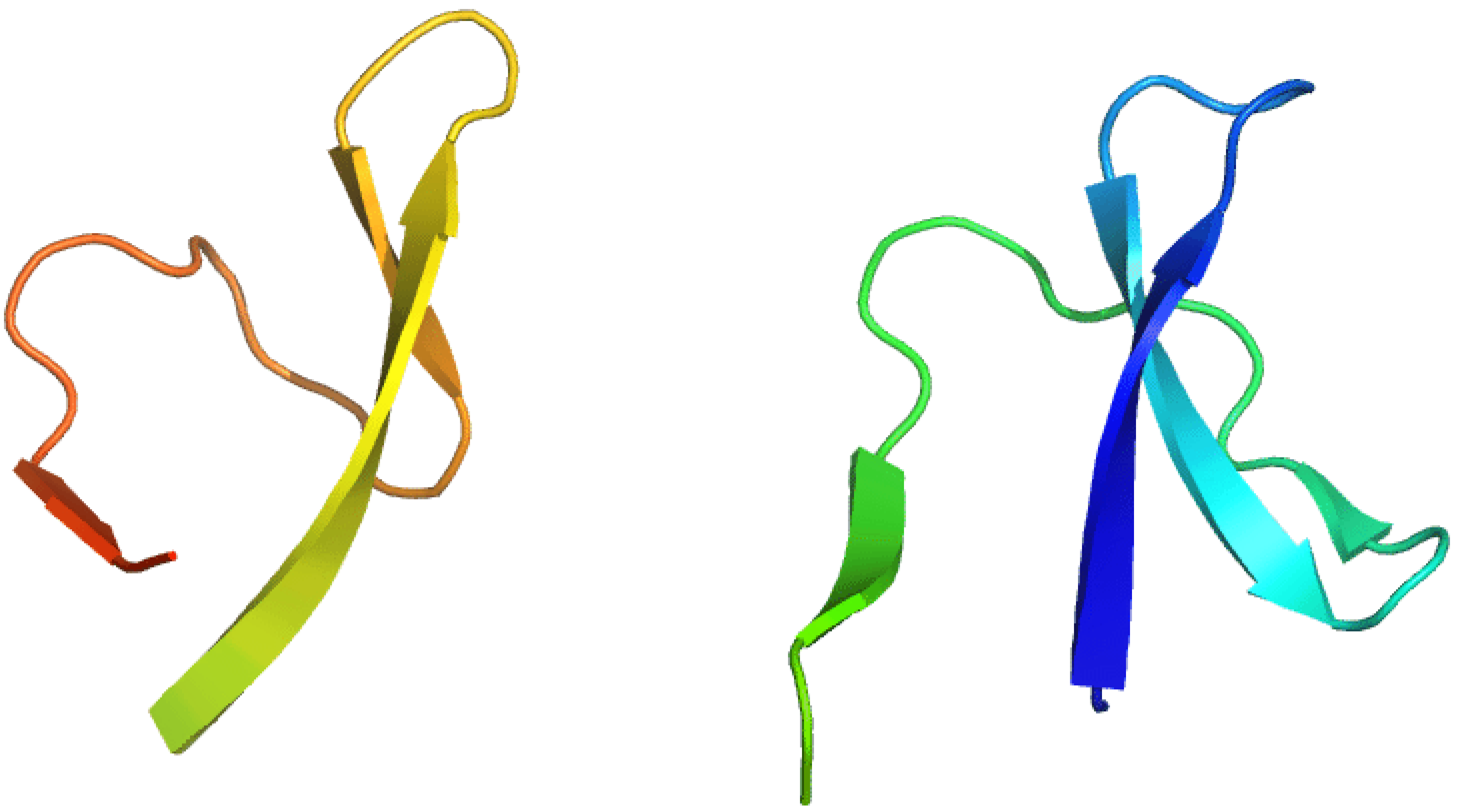
# Target 624



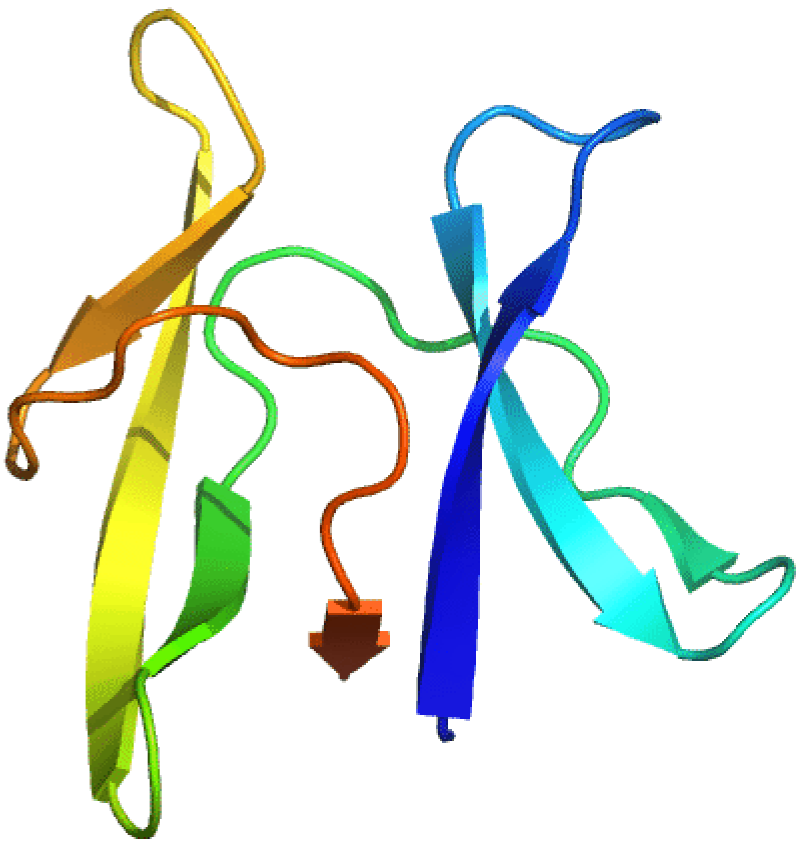
# Target 624



# Target 624

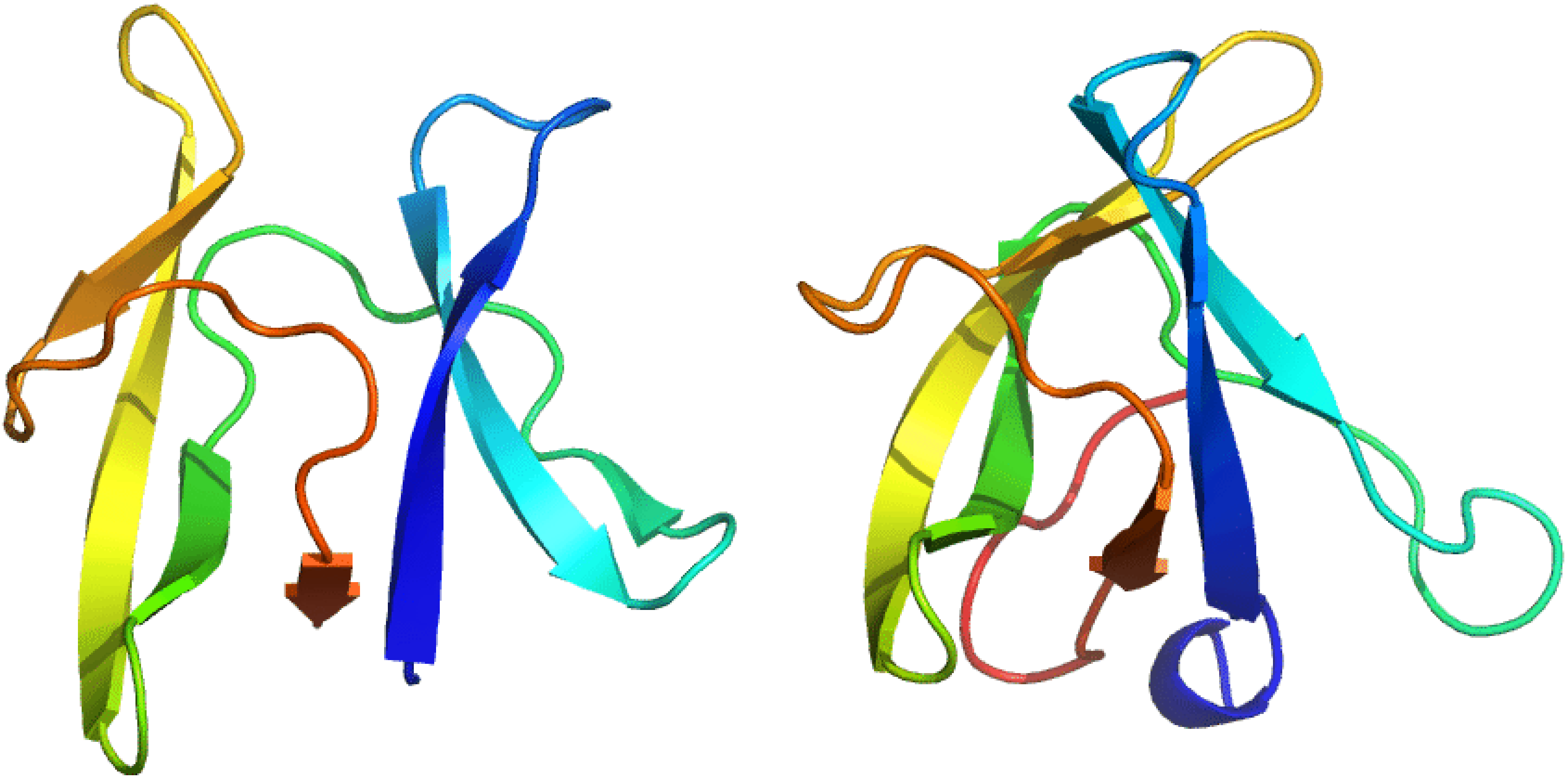


# Target 624



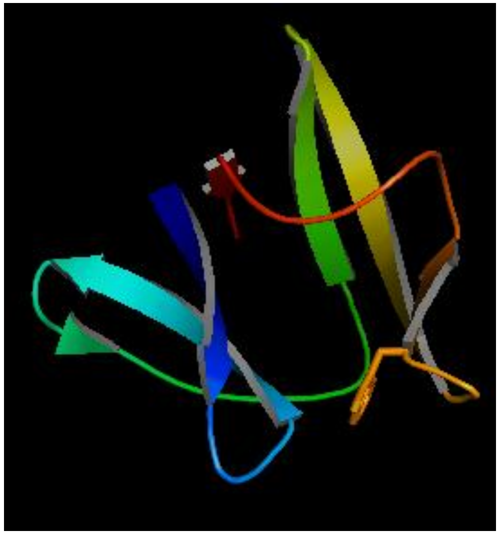
Template: 2hvy

# Target 624



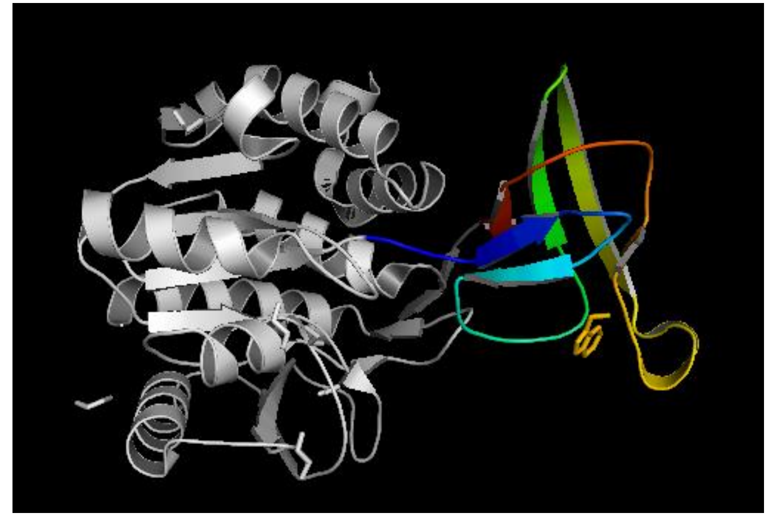
Winning model:
group **172** model **1**

# T0624: A loosened cradle-loop barrel?



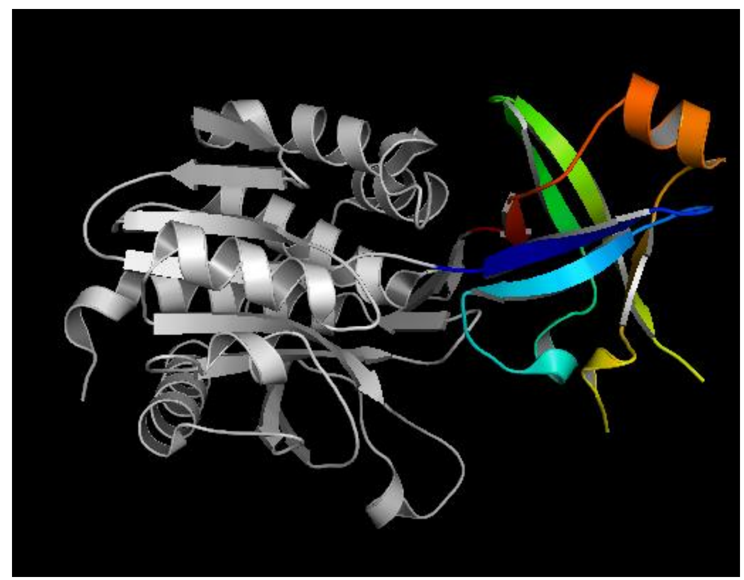
T0624

Dali z 3.5, id 17%



Putative M42 glutamyl aminopeptidase (3cpx)

Dali z 3.2, id 8%

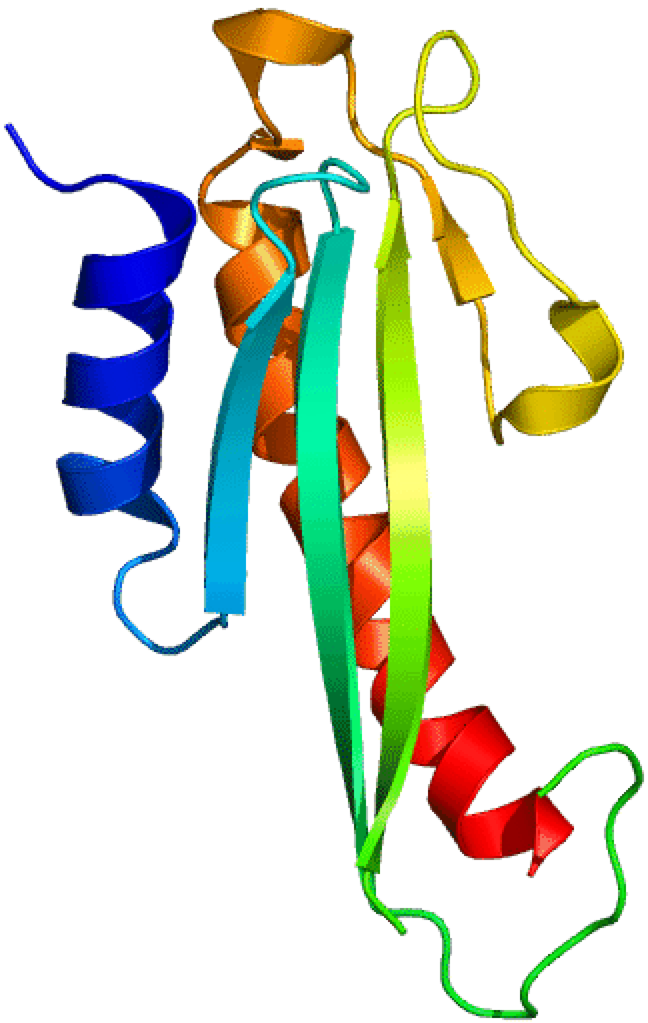


Archaeal aminopeptidase (1xfo)

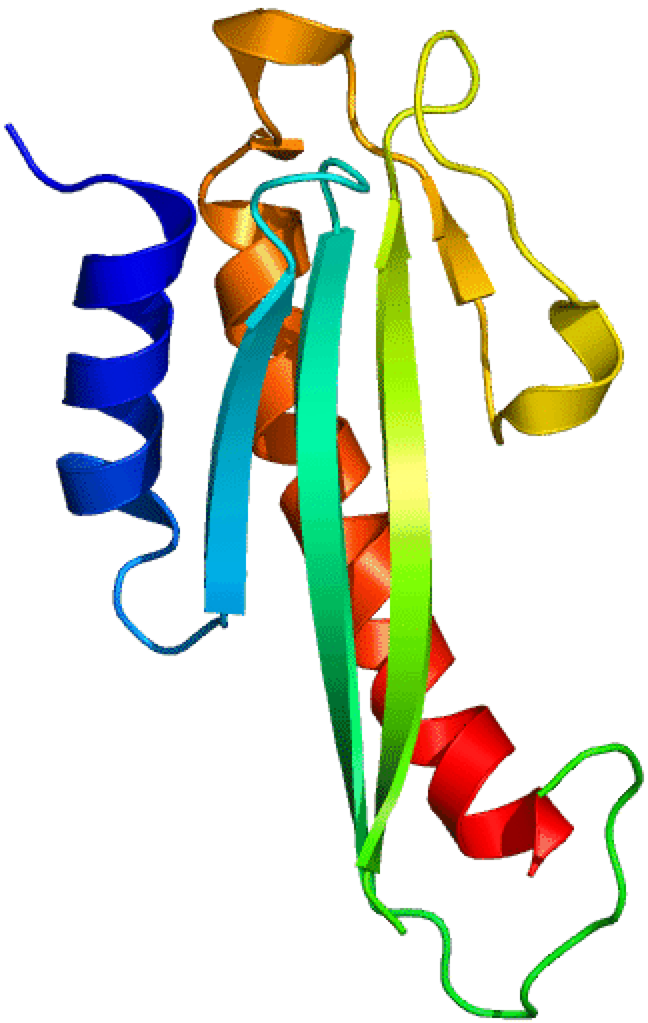
```
T0624  reGTLFYdtetgrydIRFDlesfYGGLHCGECFDVKVKDVWVP
3cpx   igFTVSY----nnhlHPIG----SPSAKEGYRLVGKDSNGDIE

T0624  VRIEXGD-DWYLVGLNVsrlDGLRVRX
3cpx   GVLKIVDeEWXLETDRL-idRGTEVTF
```

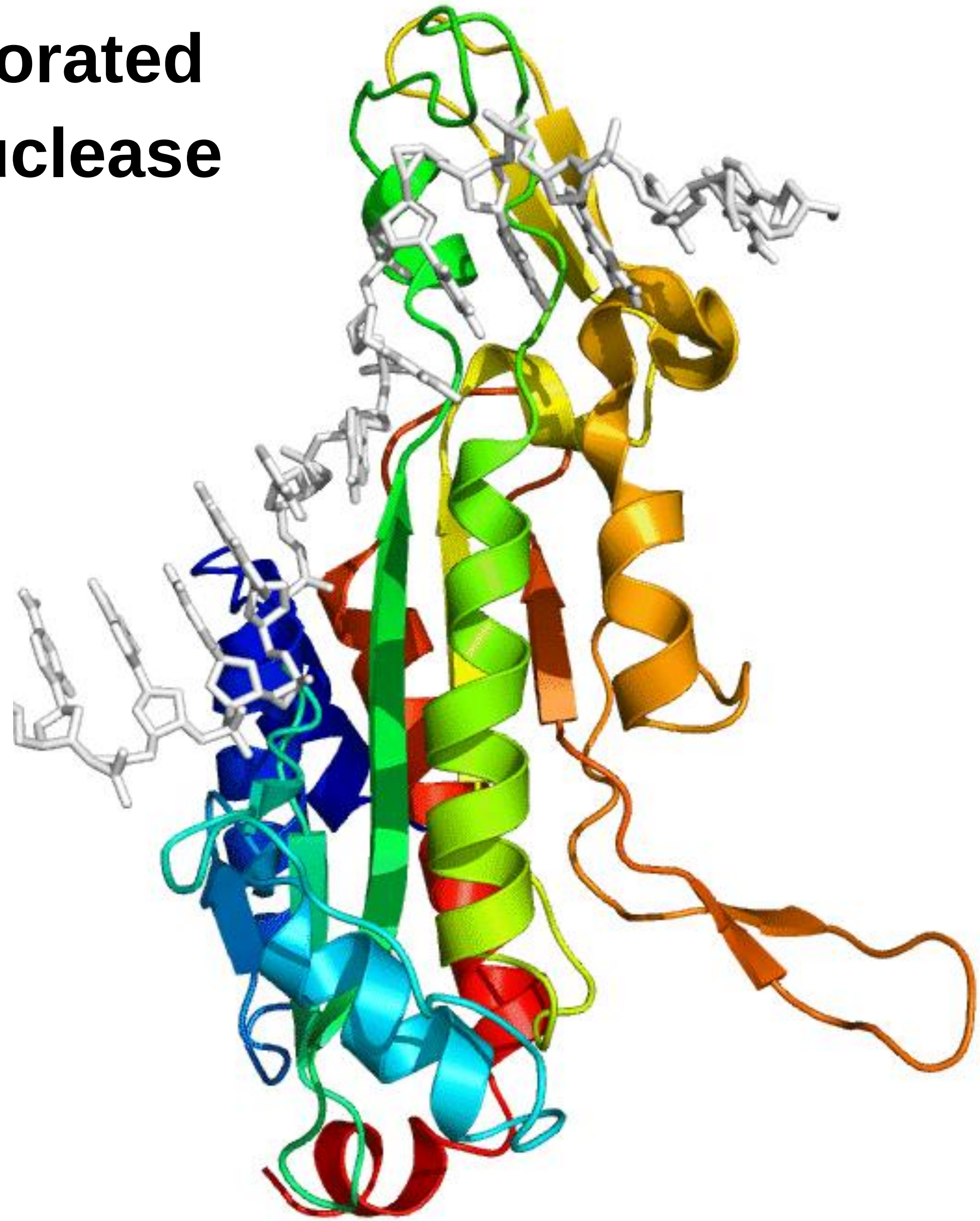
The first Dali hit is 3cpx. 3cpx and 1xfo are homologous, since they are both aminopeptidases and they have the same domain architecture (one Rossmann domain and one barrel insertion). Compared to 3cpx and 1xfo, the first two strands in T0624 are somewhat peeled off.

# Target 578



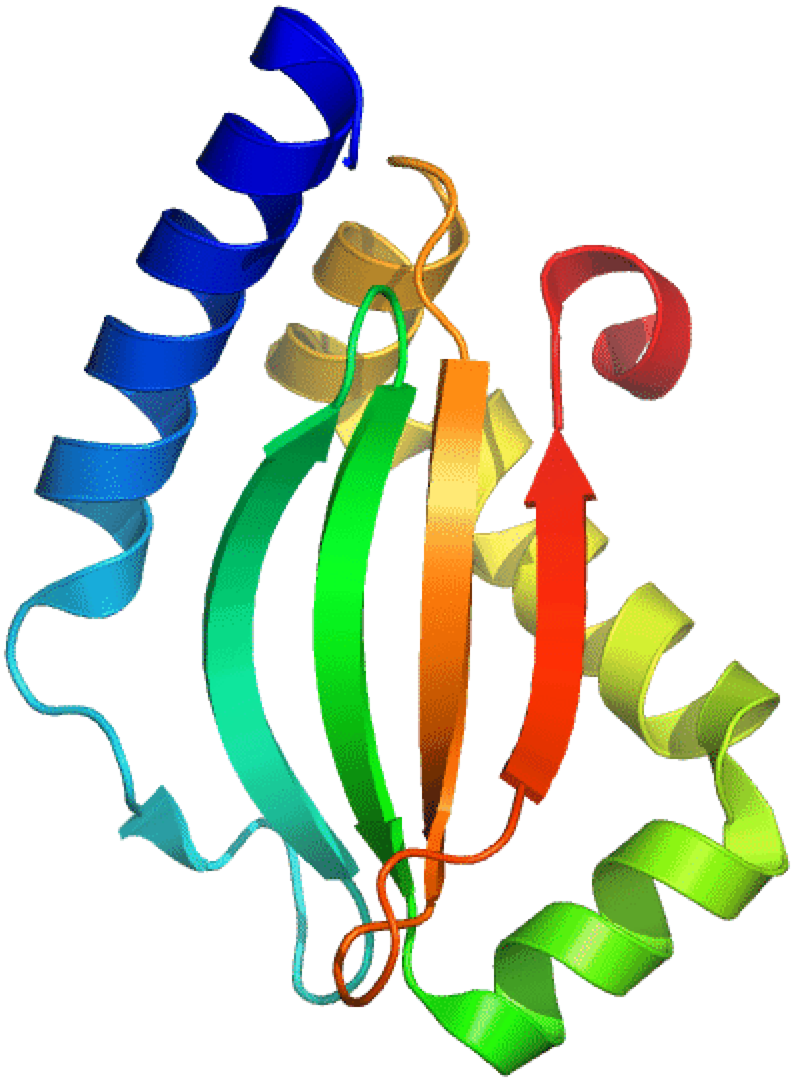
# T0578 is a deteriorated restriction endonuclease

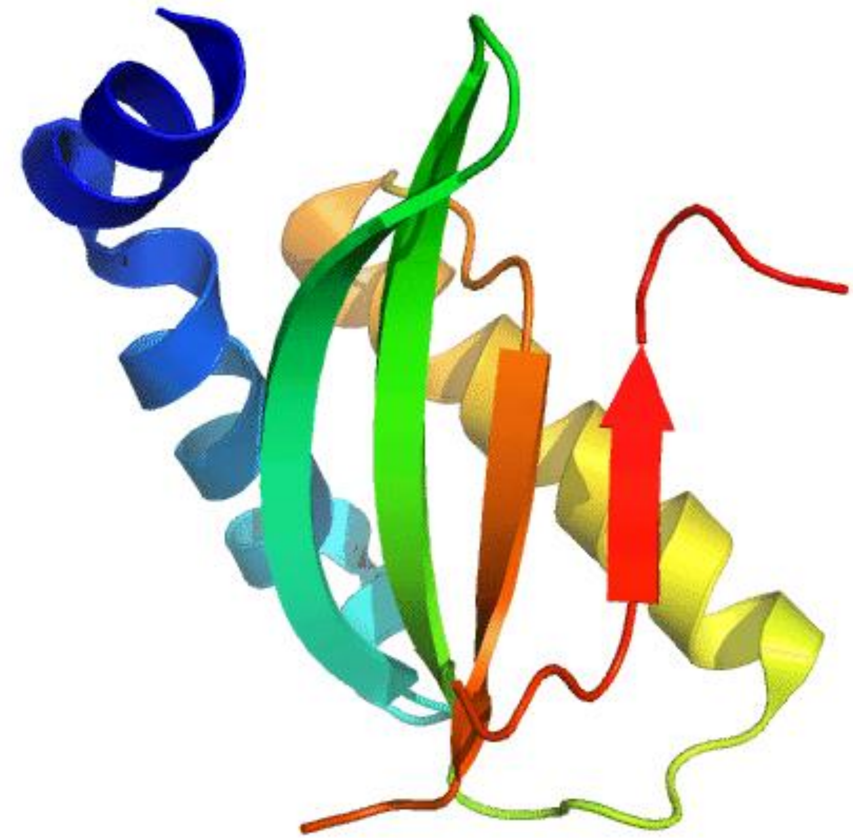




EcoRI restrictase: 1qrh

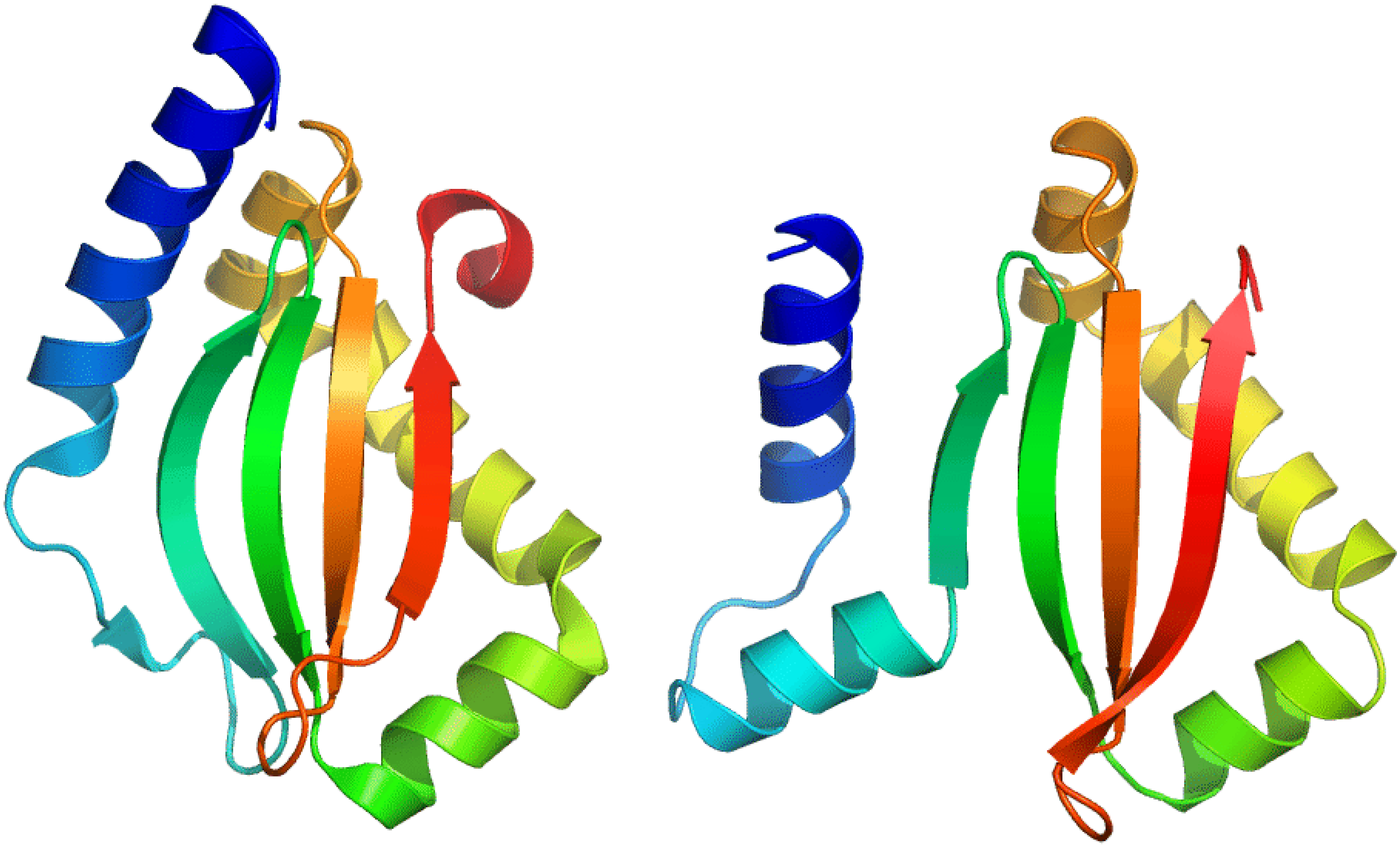
# Target 581





fatty acyl-adenylate ligase
C-terminal domain : 3lnv

# Target 581

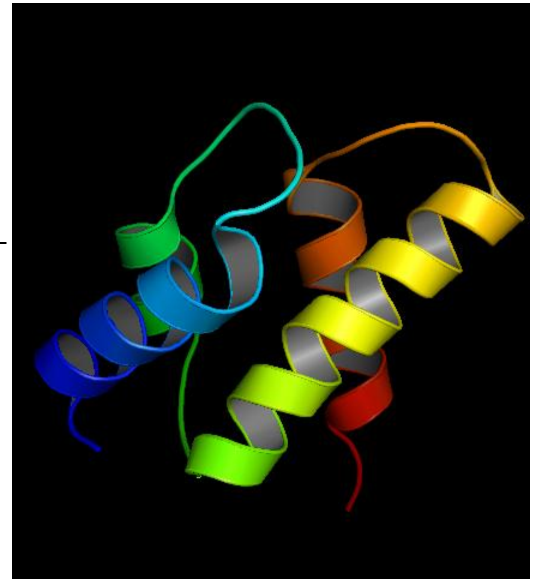


Winning model:
group **321** model **4**

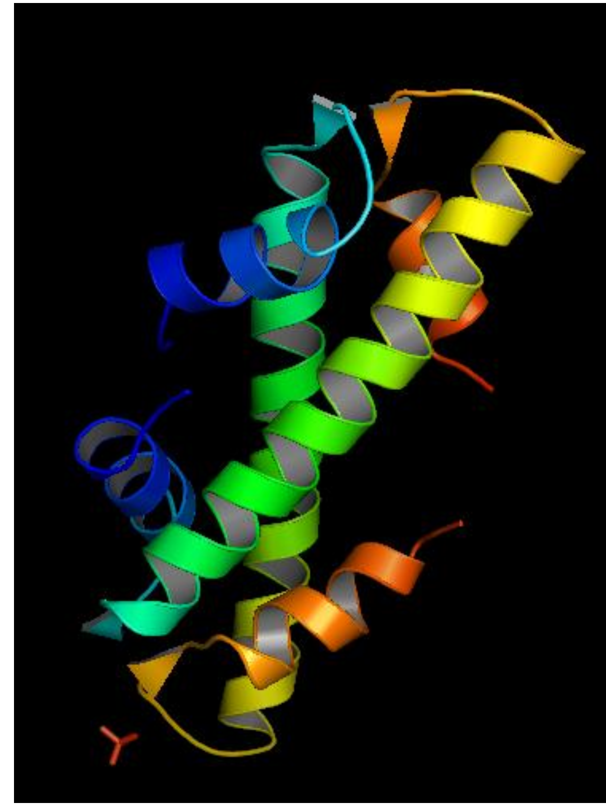
# T0538: a truncated histone fold?

T0538

Dali z
5.8



ATP-dependent protease FtsH
C-terminal domain (1lv7)



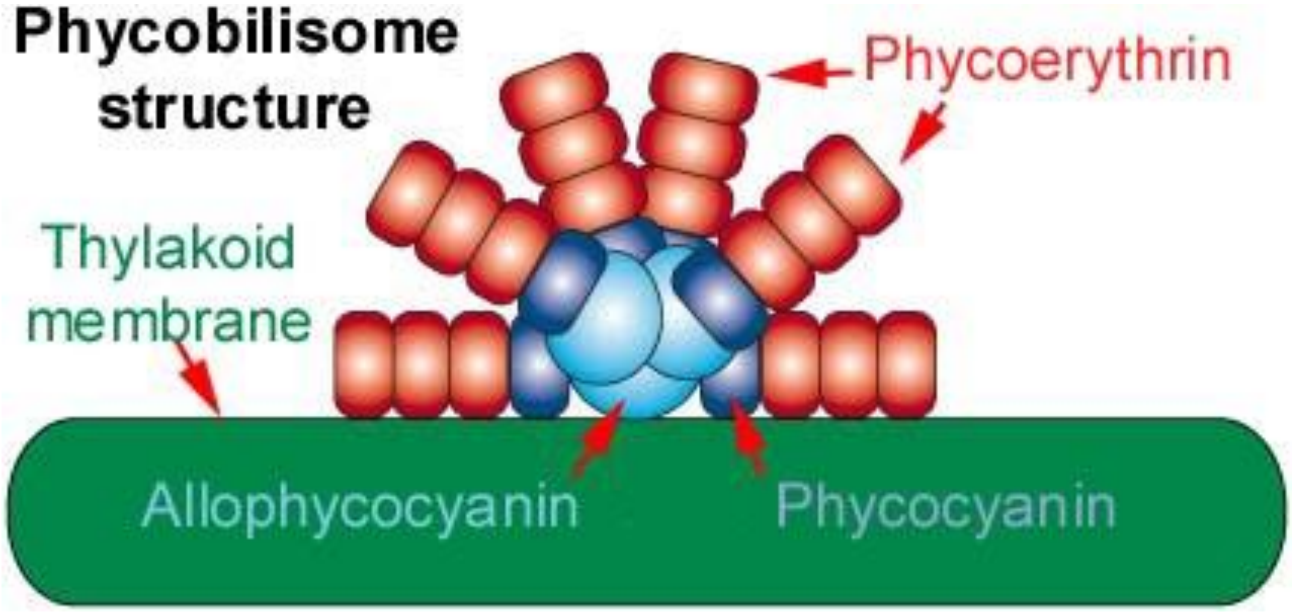
Archaeal histone (1b67)

```
T0538   MNLRWTSEAKT-kLKNIP----FFARSQAKARIEQLARQAEQDIVTPELVEQARLEFGQLE
1lv7    RRVPLAPDIDAaiIARGTpgfsGADLANLVNEAALFAARGNKRVVSMVEFEKAKDKIMMGL
```

Reference “On the origin of the histone fold” suggests homology between extended AAA-ATPase C-terminal domains and histones. T0538 lacks the first helix in the 4-helical bundle. BLAST shows that many homologs do have extra N-terminal residues and some homologs are annotated as ‘proto-chlorophyllide reductase 57 kD subunit’.

# T0544, T0553, T0554 (cancelled) and T0555

**Similar sequences from Pfam family PBS_linker_poly (PF00427):**
**Phycobilisome linker polypeptide**

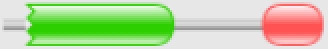



Phycobilisome: light harvesting complex of Cyanobacteria

http://www.botany.hawaii.edu/faculty/webb/BOT311/Rhodophyta/PhycobilisomeDraw.jpg

**There are 148 sequences with the following architecture: PBS_linker_poly, CpcD**

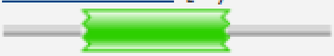
PYR1_ANASP [Anabaena sp. (strain PCC 7120)] Phycobilisome 32.1 kDa linker polypeptide, phycocyanin-associate



Show all sequences with this architecture.

**There are 129 sequences with the following architecture: PBS_linker_poly**

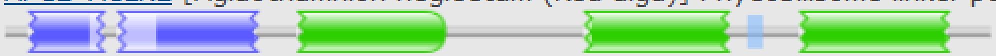
PHEG_SYNPY [Synechococcus sp. (strain WH8020)] Phycoerythrin class 2 subunit gamma, linker polypeptide (293



Show all sequences with this architecture.

**There are 37 sequences with the following architecture: Phycobilisome x 2, PBS_linker_poly x 3**

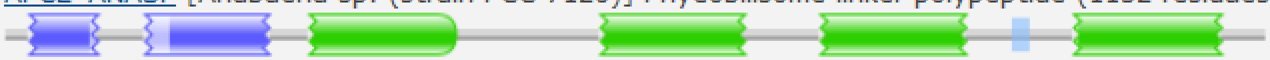
APCE_AGLNE [Aglaothamnion neglectum (Red alga)] Phycobilisome linker polypeptide (885 residues)



Show all sequences with this architecture.

**There are 12 sequences with the following architecture: Phycobilisome x 2, PBS_linker_poly x 4**

APCE_ANASP [Anabaena sp. (strain PCC 7120)] Phycobilisome linker polypeptide (1132 residues)



Show all sequences with this architecture.

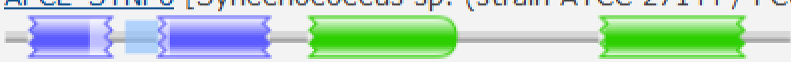
**There are 9 sequences with the following architecture: PBS_linker_poly x 2, CpcD**

Q05Q40_9SYNE [Synechococcus sp. RS9916] Phycobilisome linker polypeptide (548 residues)

Show all sequences with this architecture.

**There are 4 sequences with the following architecture: Phycobilisome x 2, PBS_linker_poly x 2**

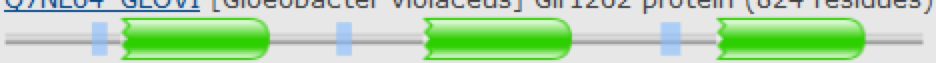
APCE_SYNP6 [Synechococcus sp. (strain ATCC 27144 / PCC 6301 / SAUG 1402/1) (Anacystis nidulans)] Phycobilis



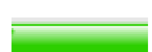
Show all sequences with this architecture.

**There are 2 sequences with the following architecture: PBS_linker_poly x 3**

Q7NL64_GLOVI [Gloeobacter violaceus] Glr1262 protein (824 residues)

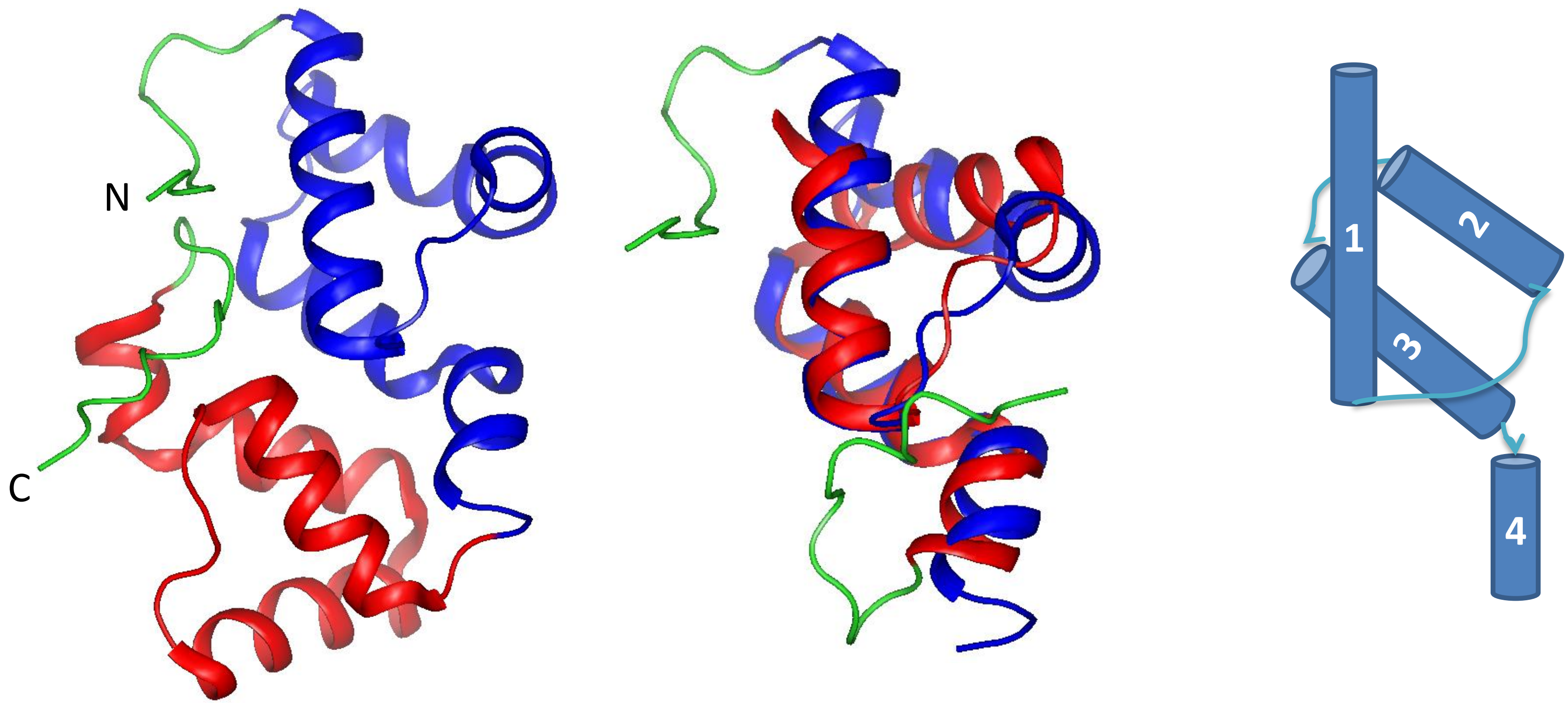


Show all sequences with this architecture.

 **: PBS_linker_poly domain**

http://pfam.sanger.ac.uk/family?acc=PF00427

# PBS_linker_poly itself is a duplication consisting of two helical domains




```
                                H1                    H2                    H3          H4
T0553_domain1  -MKVFKRVAGIKDKAAIKTLISAAYRQIFERDIAPYIAQNEFSGWESKLGNGEITVKEFIEGLGYSNLYLKEFYTPY----------
T0553_domain2  ---------------PNTKVIELGTKHFLGRAP---IDQAEIRKYNQILAT--QGIRAFINALVNSQEYNEVFGEDTVPYRRFPTLE
```

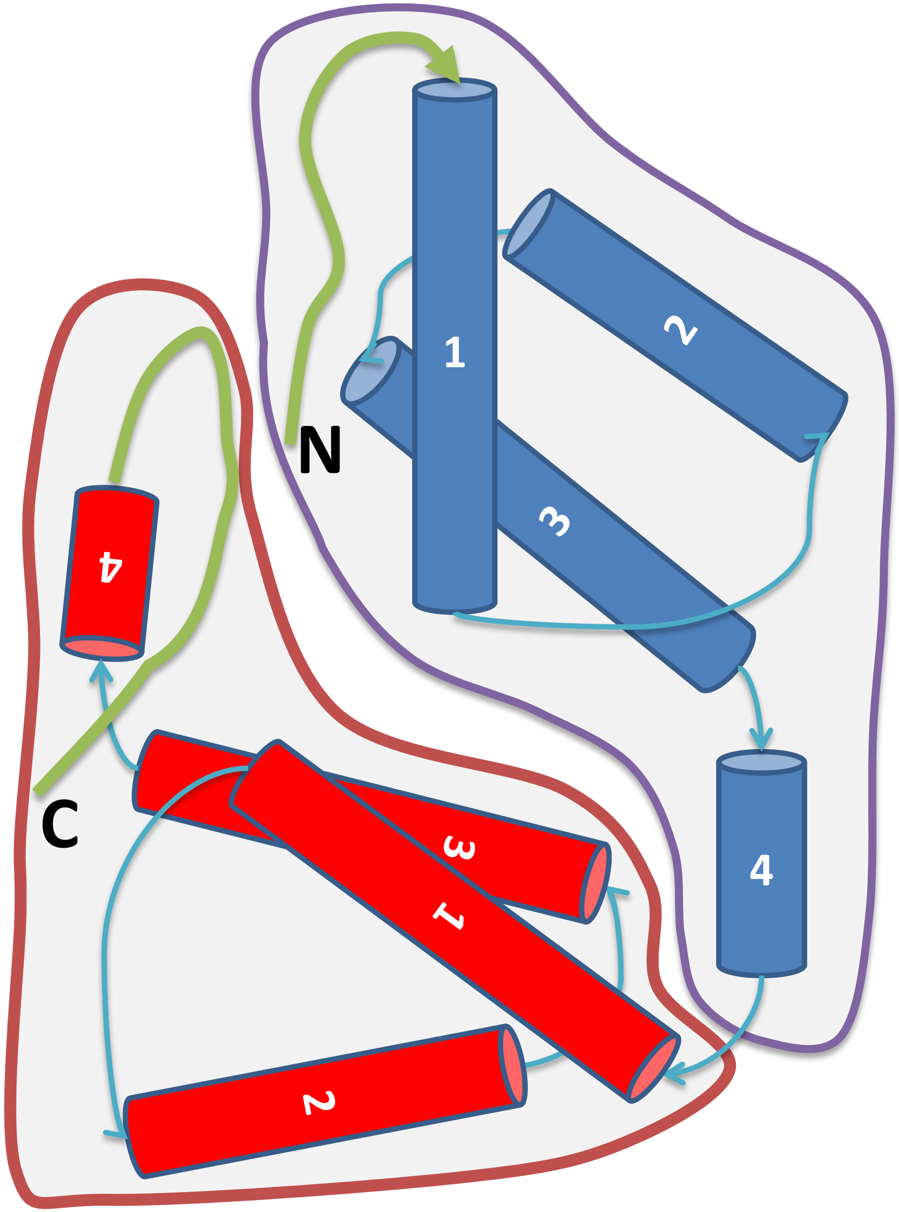
# Duplication in T0553 can be recognized by HHpred

```
>PF00427 PBS_linker_poly:  Phycobilisome Linker polypeptide
  Probab=80.37    E-value=4  Score=30.72  Aligned_cols=57  Identities=23%
Similarity=0.325  Sum_probs=0.0
```

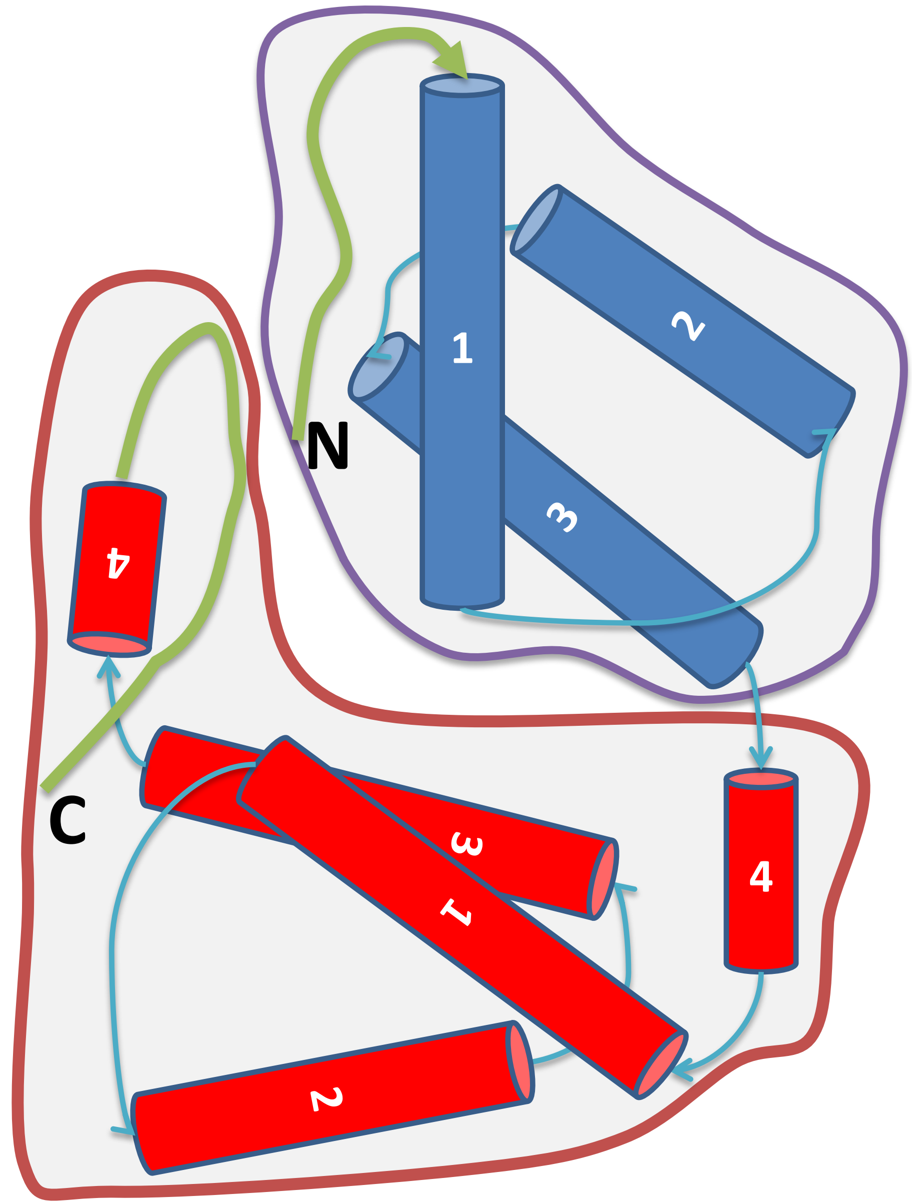
Domain 1:


```
Q ss_pred                HHHHHHHHHHHhCCCcchhhhhccchHHHHHHHhcCCccHHHHHHHHHcCHHHHHHhcccCCcc
Q Tue_Nov_30_18:      18 TLISAAYRQIFERDIAPYIAQNEFSGWESKLGNGEITVKEFIEGLGYSNLYLKEFYTPYPNT   79 (141)
Q Consensus           18 ~vI~AaYrQVf~~~~~~~~~~~~~rl~~lESqLr~g~IsVreFVr~LakS~~yr~~f~~~~~~~   79 (141)
                        .+|..+++.++||   ++....+...+=.-+-..-   ...||..|.-|+.|.+.|=+..-||
T Consensus           72 R~iEl~~khlLGR---ap~~~~Ei~~~~~i~a~~G--~~a~Id~lldS~EY~~~FG~d~VPy  128 (131)
T PF00427_consen      72 RFIELNFKHLLGR---APYNQAEISAYSIILAEKG--FEAFIDSLLDSDEYLENFGEDTVPY  128 (131)
T ss_pred                HHHHHHHHHHHhCC---CCCCHHHHHHHHHHHHHhcC--hHHHHHHHhCCHHHHHHcCCCCCCC
```

Domain 2:



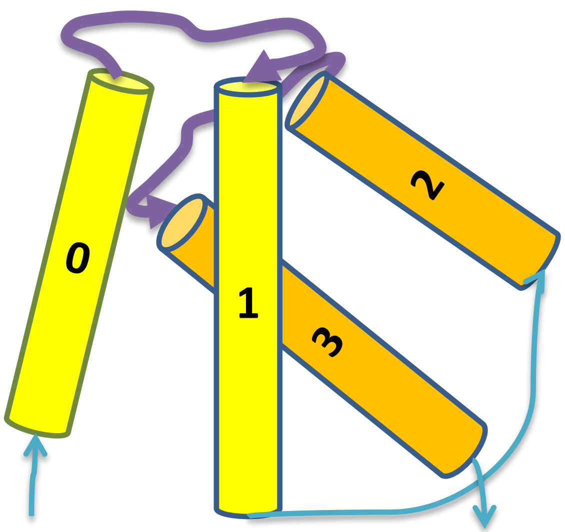

Domain definition according to sequence.



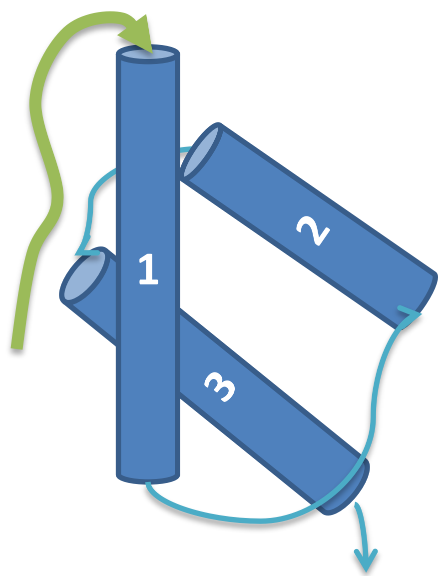

Domain definition according to structure.

# Using EF-hand structures as templates

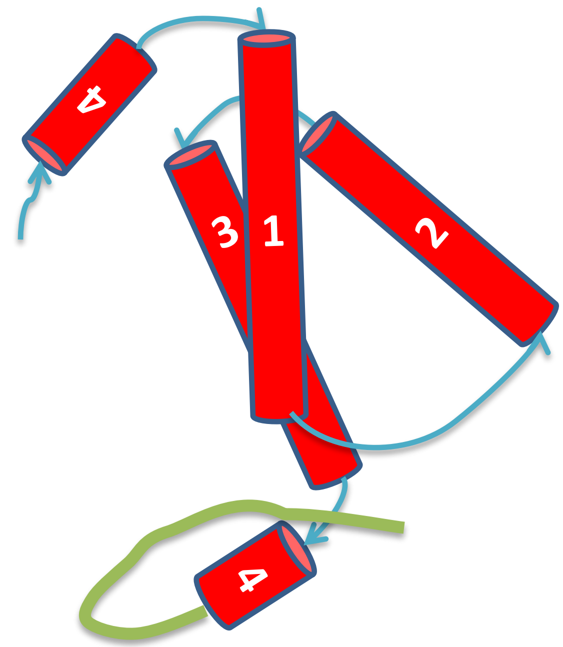
```
>1snl_A Nucleobindin 1, calnuc; EF-hand, calcium-binding, metal binding protein; NMR {Homo sapiens} SCOP:
  a.39.1.7
  Probab=65.11   E-value=6.4   Score=26.97   Aligned_cols=56   Identities=9%   Similarity=0.150   Sum_probs=0.0

  Q ss_pred                 CHHHHHHHHHHHHHHHHhCCCcchhhhccchHHHHHHh----------cCCccHHHHHHHHHcCHH
  Q Tue_Nov_30_18:     12 DKAAIKTLISAAYRQIFERDIAPYIAQNEFSGWESKLG----------NGEITVKEFIEGLGYSNL   67 (141)
  Q Consensus          12 ~~~~le~vI~AaYrQVf~~~~~~~~~~~~rl~~lESqLr----------~g~IsVreFVr~LakS~~   67 (141)
                            |..++..++.+...++.+...........-..++..+.          +|.||.-||++++.+.++

  T Consensus          38 ~~~El~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~D~d~DG~Is~eEF~~~~~k~ef  103 (103)
  T 1snl_A             38 DEQELEALFTKELEKVYDPKNEEDDMREMEEERLRMREHVMKNVDTNQDRLVTLEEFLASTQRKEF  103 (103)
  T ss_dssp               EHHHHHHHHHHHHHHTTSCCSSCSSHHHHTTHHHHHHHHHHHHHTCSSCSSEEEHHHHHHHHHCCCC
  T ss_pred               CHHHHHHHHHHHHHHHhccchhhhhhhhhhHHHHHHHHHHHHHHHhCCCCCCcCcHHHHHHHHhccCC
```




Two EF-hands

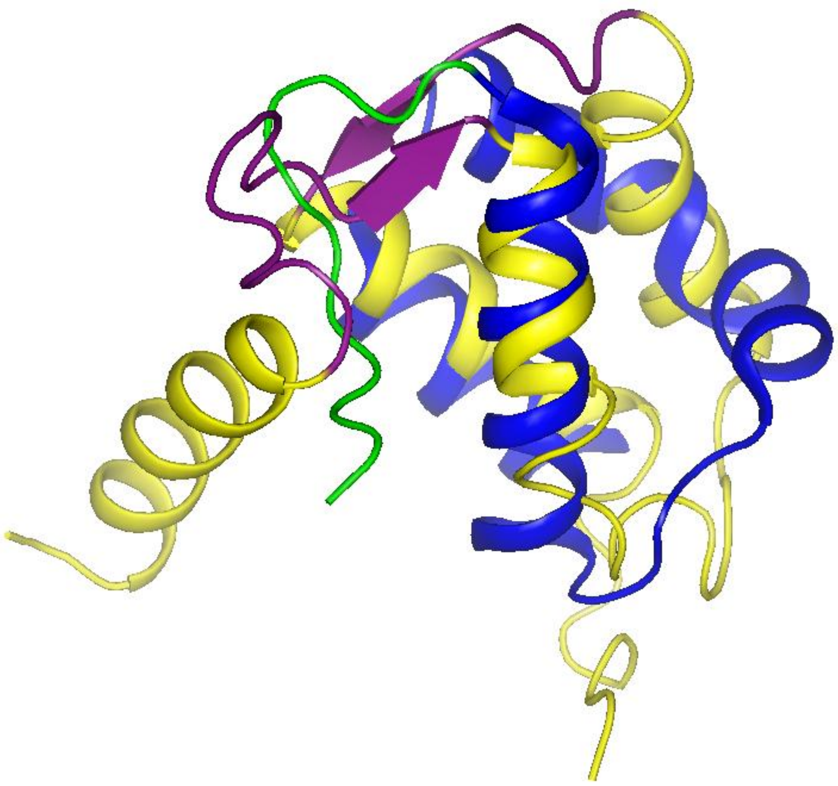



T0553_domain1

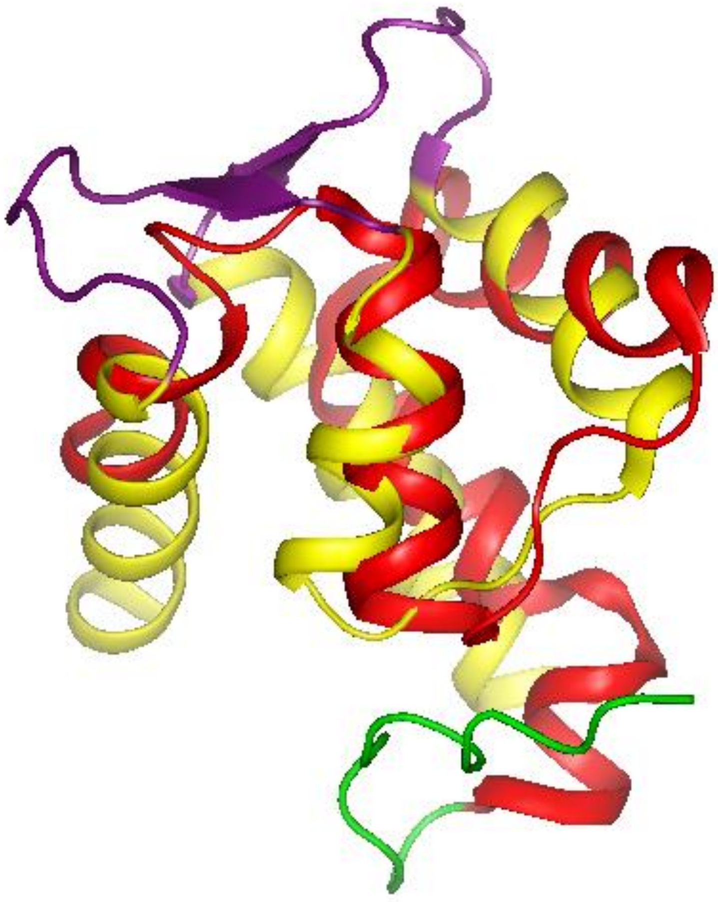



T0553_domain2

**T0553 domain 1** aligned with a **EF-hand protein 1K2H**



**T0553 domain 2** aligned with a **EF-hand protein 2OBH**



Problems of using EF-hand structures as templates:

- High structural variations.
- Not suitable for modeling the interaction and  orientation between the two duplicated domains.

# What about canceled targets?

Some were canceled because structures for them were not determined in time

For some of them no templates can be found easily by sequence, e.g. **T0642**

**T0642** was interesting, because it is a long, 387aa protein without BLAST hits, which doesn't happen that much anymore

Since no sequence homologs can be identified for it, maybe predictions can help us shed light on evolutionary origin of this protein

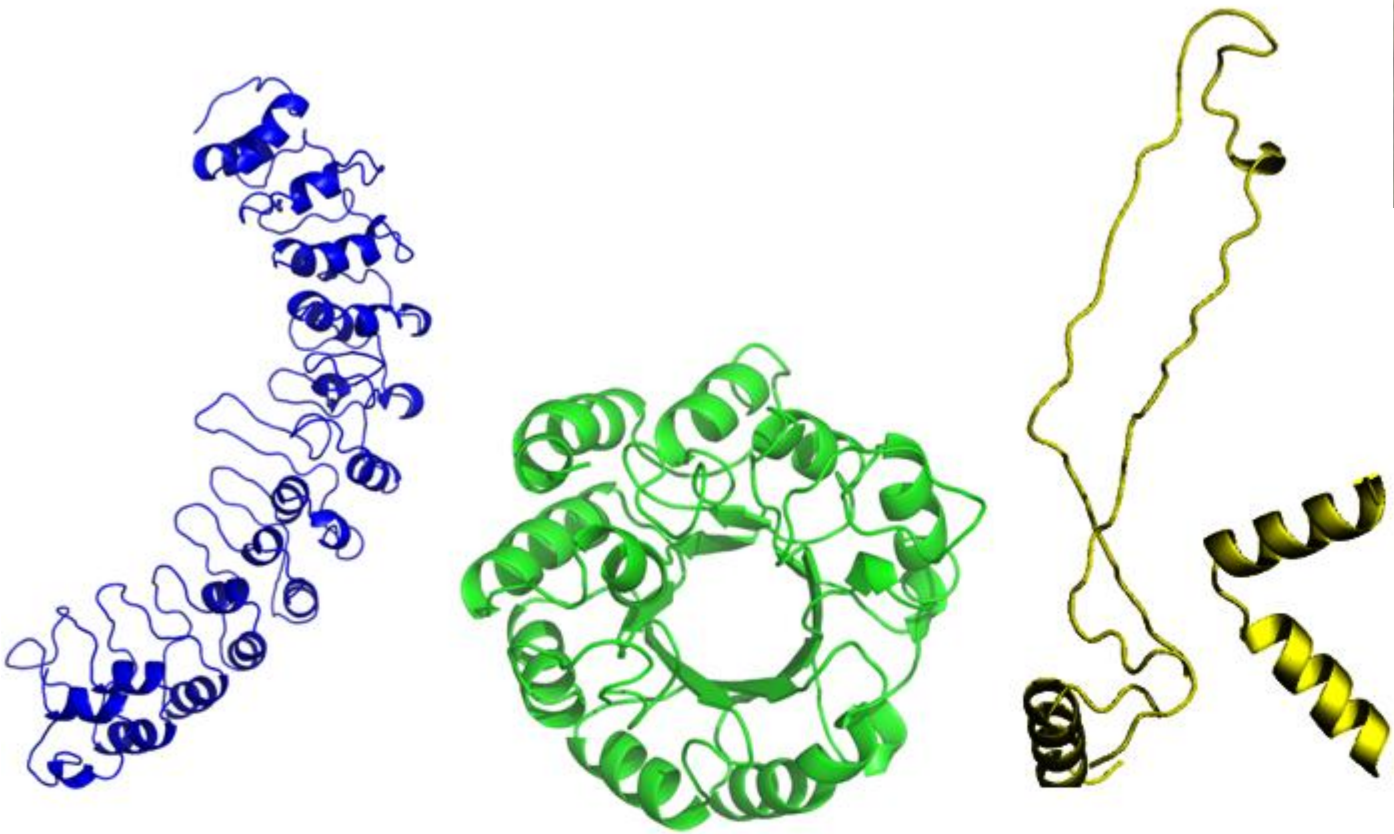
# What about canceled targets?

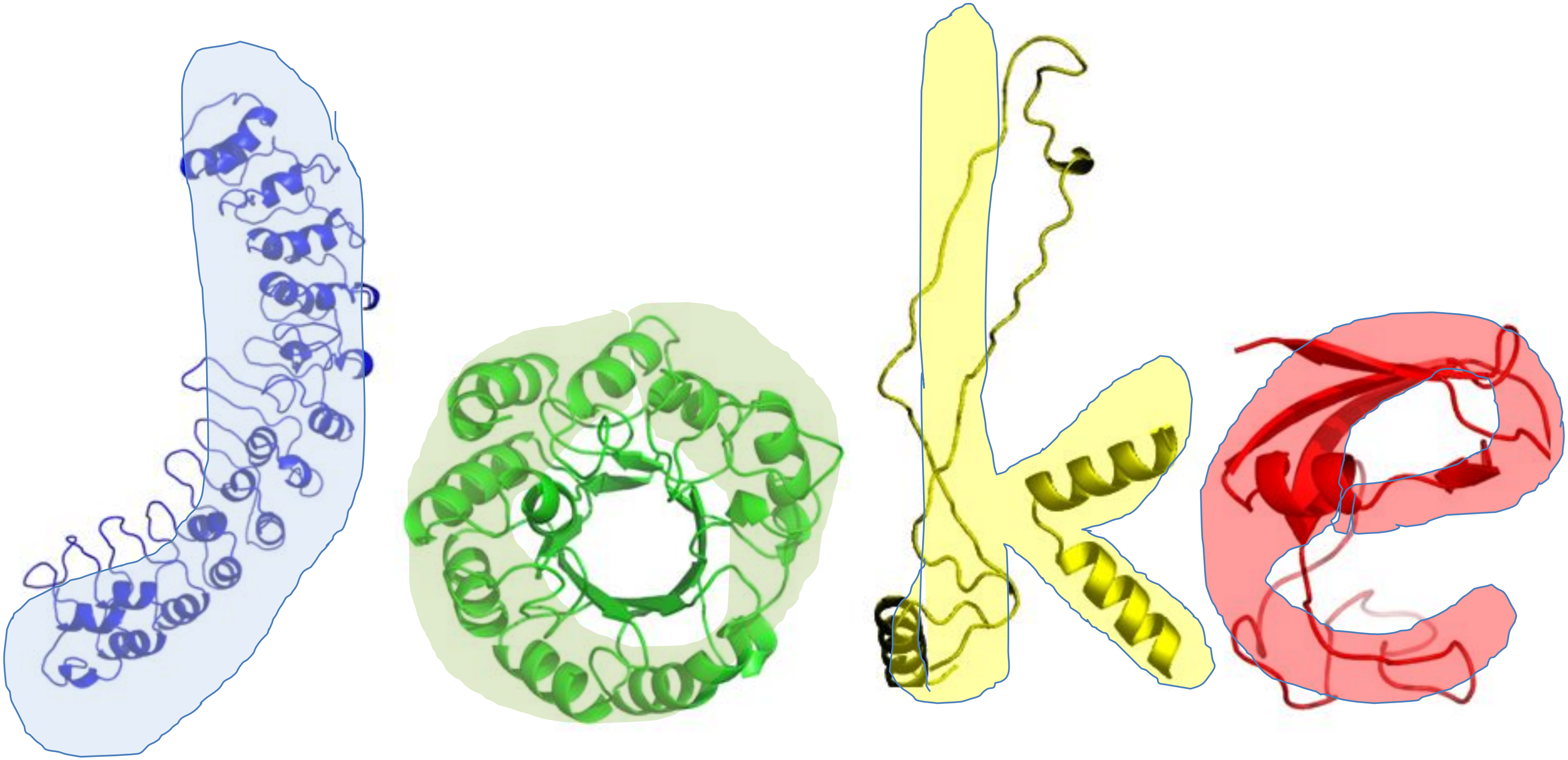
Since no sequence homologs can be identified for it, maybe predictions can help us shed light on evolutionary origin of this protein

**>T0642: J0KE1 from *Homo sapiens***
MDEARCASPERSTERRIFICNEWSYESTERDAYWERELEASEDTHELASTSEQINTHENINTHCASPPLE
ASEGETRESTEDANDLETASSESSMENTDETERMINETHEESTSCIENTIFICCENTERSTHISTARGETIS
DIFFERENTANDHASVERYSPECIFICSHAPEWILLCHECKITATTHEMEETINGINPACIFICGRVEHAHA
LASTWCFINALISTSITALYANDFRANCEWEREELIMINATEDINPRELIMINARYMATCHESSPAINWIN
AGAINSTNETHERLANDSINFINALINTERESTINGENDINGHAVEANICEFALLMERRYCHRISTMASA
NDHAPPYNEWYEARTAKEITEASYANDSMILE

# We clustered predictions, and got disparate results:

**Qian Cong**
graduate student




Joke

# Sequence analysis of 642

**>T0642: J0KE1 from *Homo sapiens***
MDEARCASPERSTERRIFICNEWSYESTERDAYWERELEASEDTHELASTSEQINTHENINTHCASPPLE
ASEGETRESTEDANDLETASSESSMENTDETERMINETHEESTSCIENTIFICCENTERSTHISTARGETIS
DIFFERENTANDHASVERYSPECIFICSHAPEWILLCHECKITATTHEMEETINGINPACIFICGRVEHAHA
LASTWCFINALISTSITALYANDFRANCEWEREELIMINATEDINPRELIMINARYMATCHESSPAINWIN
AGAINSTNETHERLANDSINFINALINTERESTINGENDINGHAVEANICEFALLMERRYCHRISTMASA
NDHAPPYNEWYEARTAKEITEASYANDSMILE

MY DEAR CASPERS, TERRIFIC NEWS: YESTERDAY WE RELEASED THE LAST SEQUENCE IN THE NINTH CASP.
PLEASE GET RESTED AND LET ASSESSMENT DETERMINE THE BEST SCIENTIFIC CENTERS.
THIS TARGET IS DIFFERENT AND HAS VERY SPECIFIC SHAPE.
WILL CHECK IT AT THE MEETING IN PACIFIC GROVE.
HAHA, LAST WORLD CUP FINALISTS, ITALY AND FRANCE WERE ELIMINATED IN PRELIMINARY MATCHES. SPAIN WIN AGAINST NETHER LANDS IN FINAL!! INTERESTING ENDING !!!
HAVE A NICE FALL☺ MERRY CHRISTMAS AND HAPPY NEW YEAR☺
TAKE IT EASY AND SMILE☺

# Talk plan

- Target Overview
- Domain Definition
- Domain Classification
- CASP9 categories: TBM and FM

# Defining CASP9 categories: TBM and FM

**TBM** assumes presence of template(s) by definition

Does **FM** assume absence of template(s) by definition?

If so, it should be called not-TBM (or TBM-not)
**but it is not!**

Presence/absence of templates is shaky ground:
some say there are templates for everything;
some say templates need to be found by sequence;
some say templates need to be found by structure.

Which method should be used for template identification?

# Defining CASP9 categories: TBM and FM

What is the difference between **TBM** and **FM**?

- clearly, templates have something to do with it;

- traditionally, predictors thought about FM as “hard”;

FM, which is “free modeling”,
**a category where predictors are free to do whatever they can, they can’t get it right ANYWAY**

# Listen to your data!

**Cutoffs, changes, strategies should come naturally from the data you have**

# Idea:

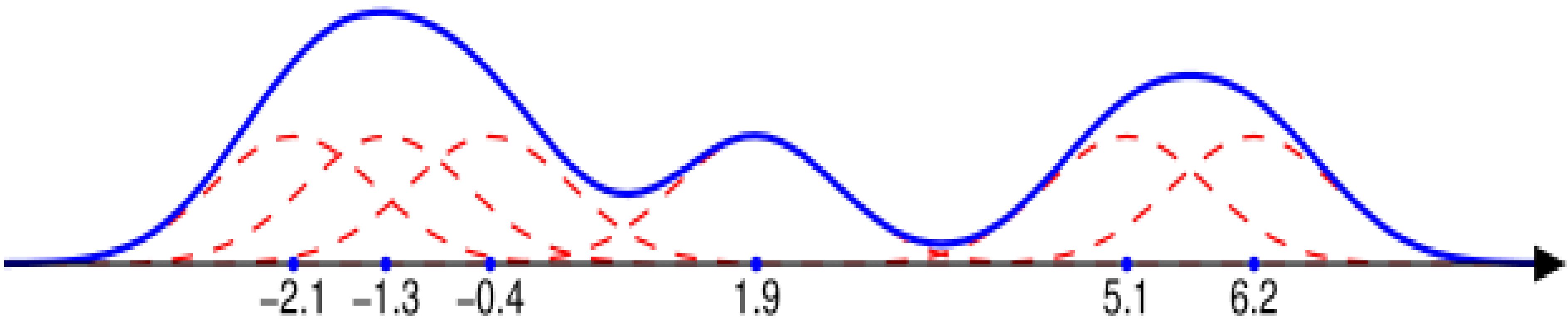
1) categories should depend on predictions and

2) boundaries between categories should come out naturally from the data

# Let’s see what predictions tell us

**Gaussian kernel density estimation!**

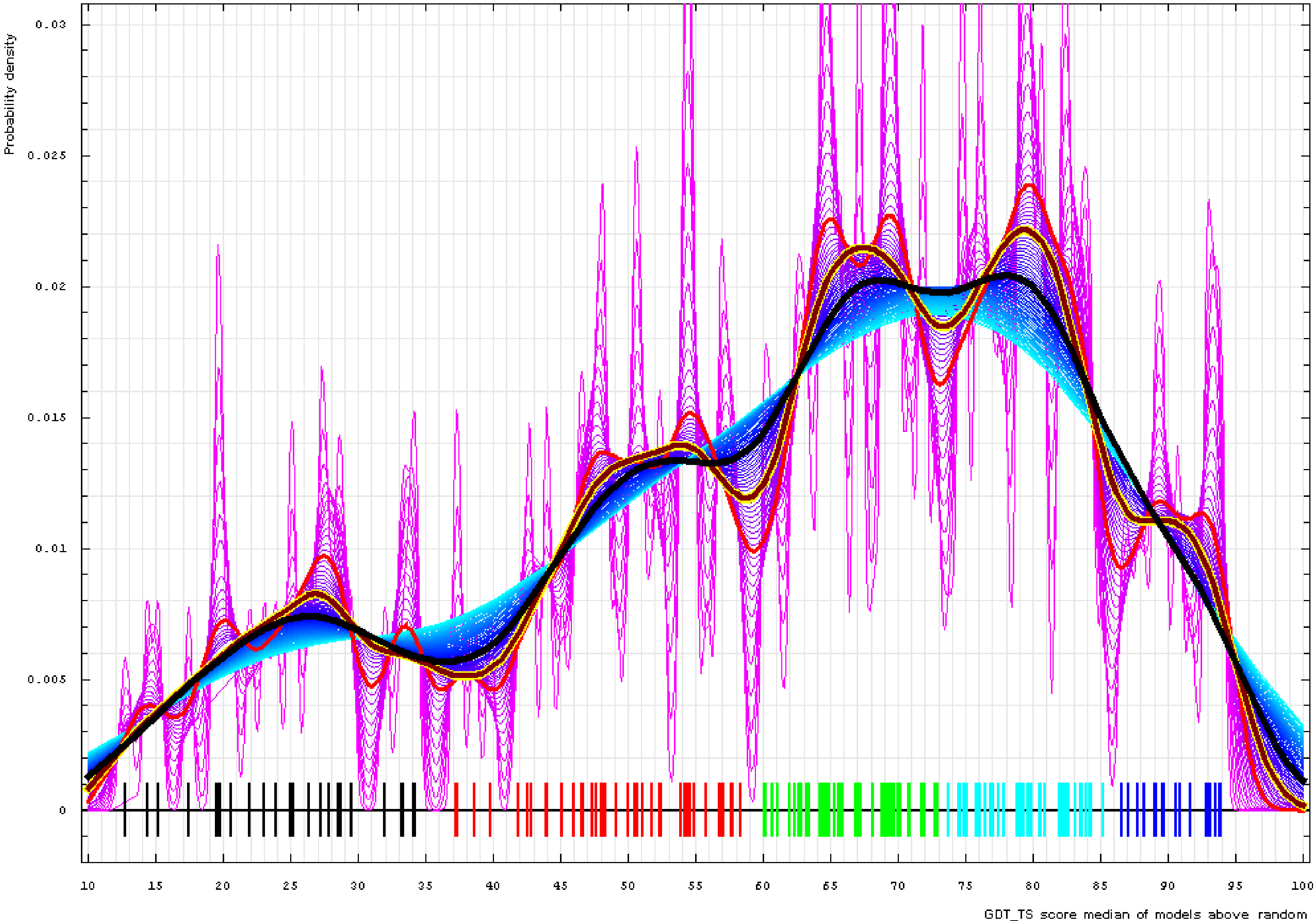
$$\widehat{f}_h(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) \qquad K(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}.$$




http://en.wikipedia.org/wiki/Kernel_density_estimation

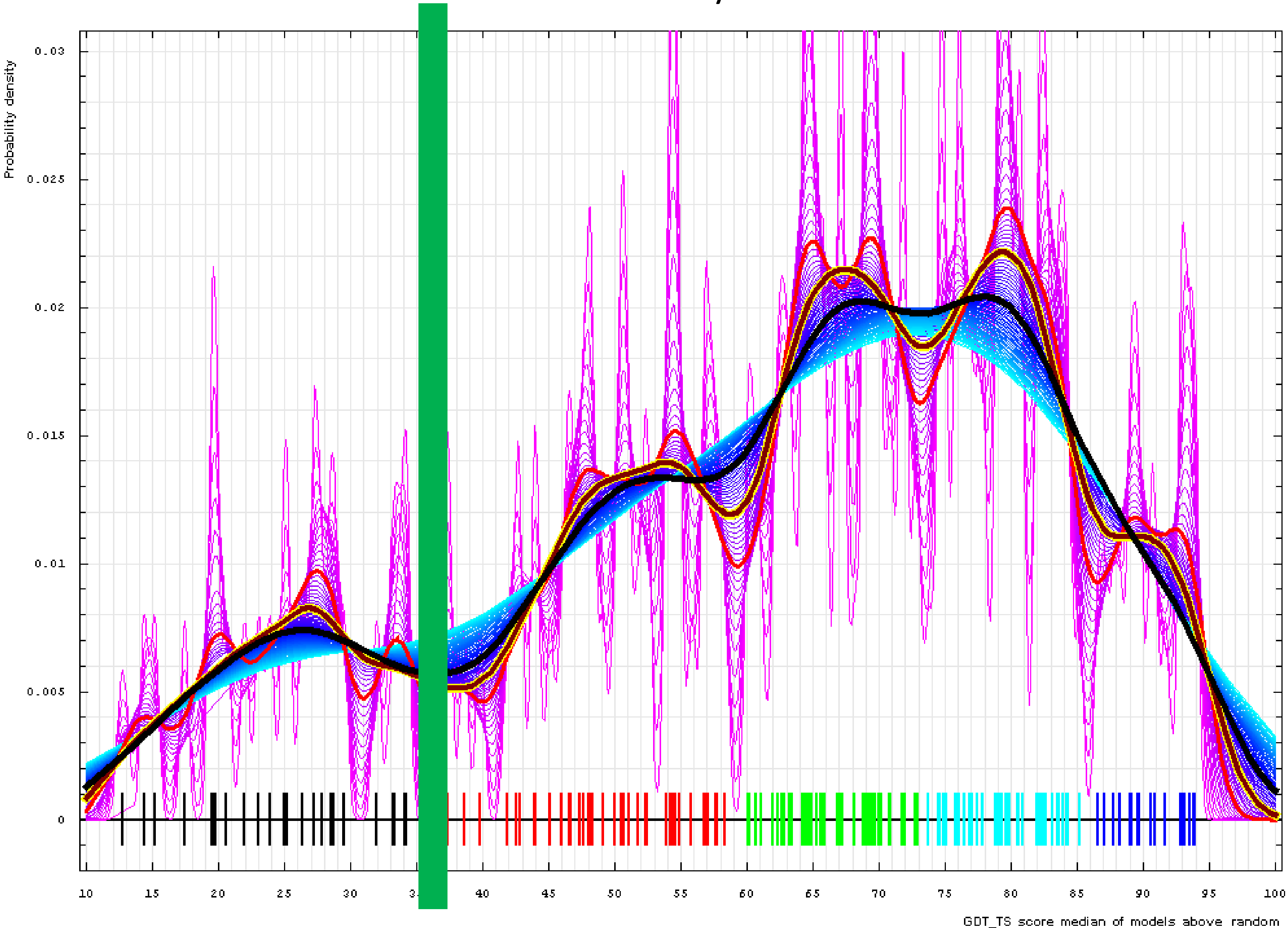
# Median GDT_TS for above random models

## Gaussian Kernel density estimation



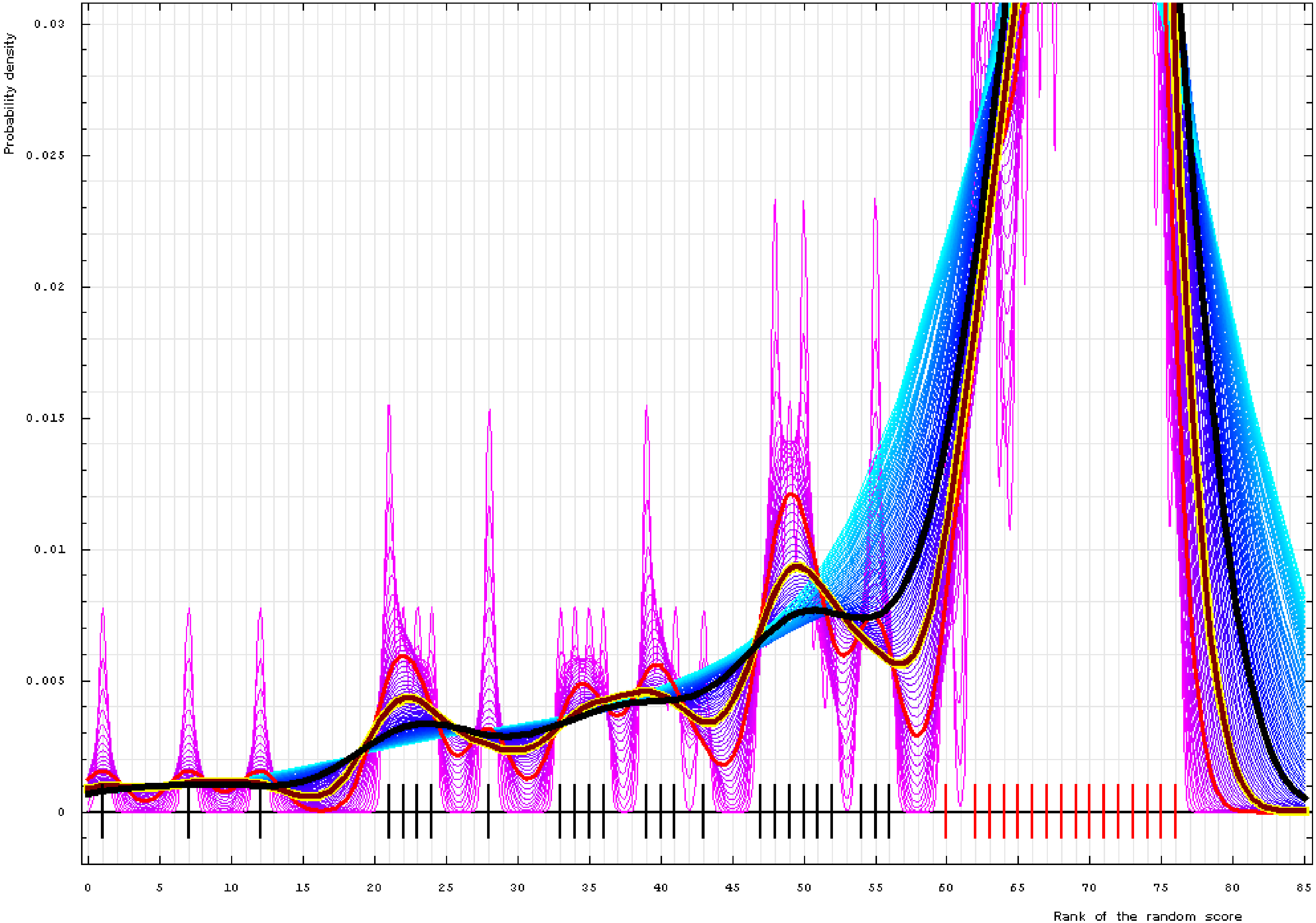
# Median GDT_TS for above random models

Gaussian Kernel density estimation



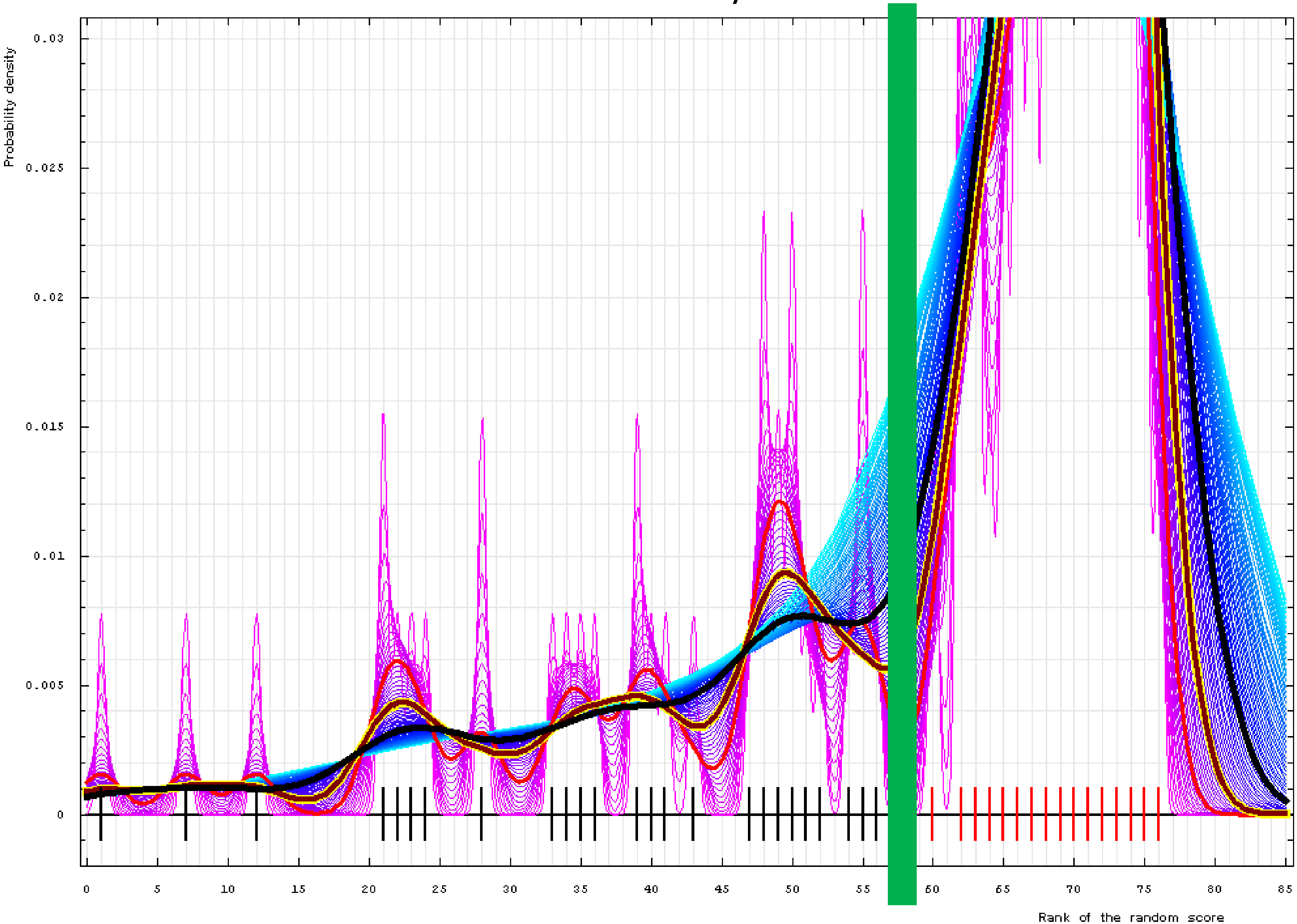
# Rank of the random model

## Gaussian Kernel density estimation



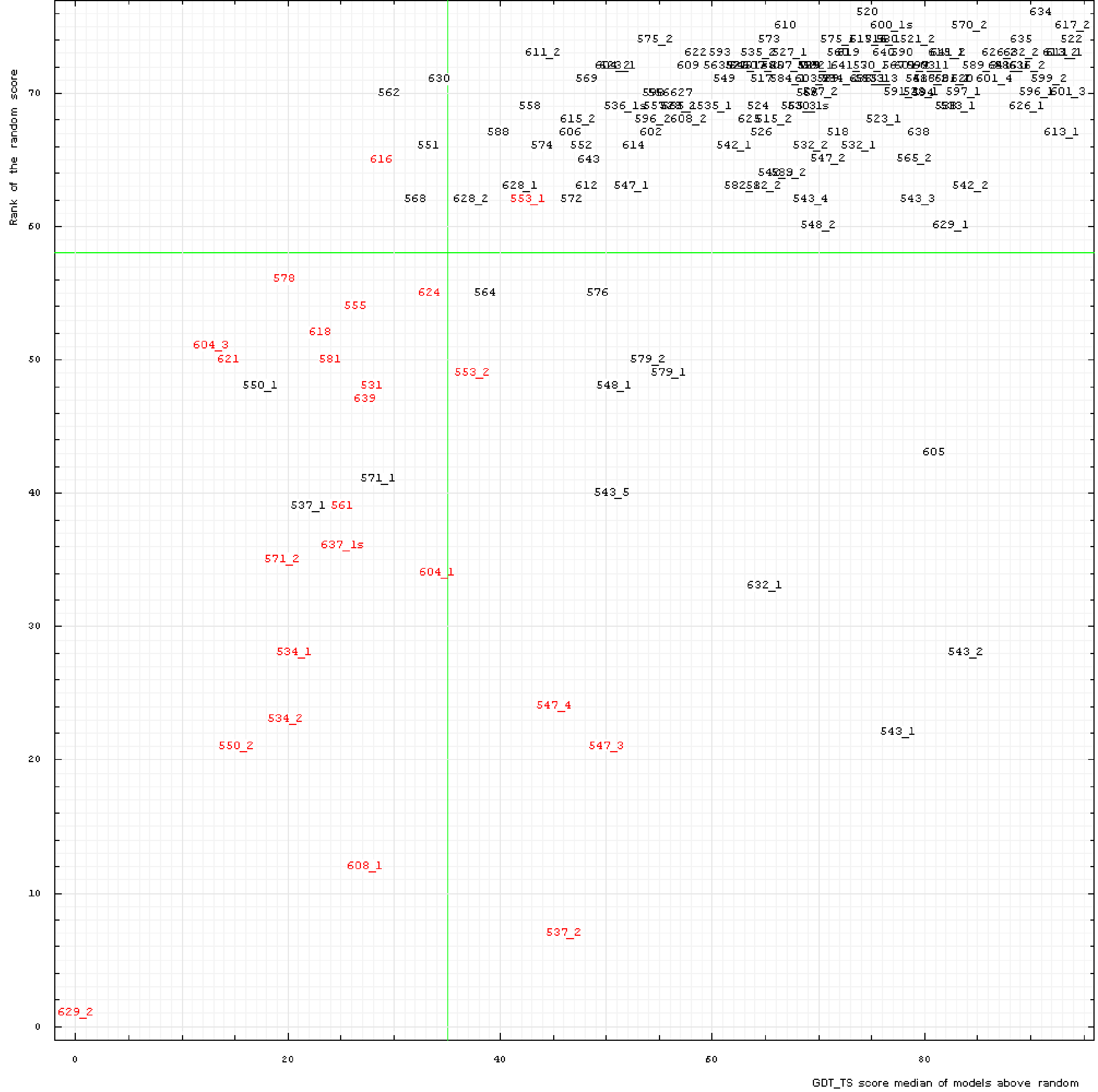
# Rank of the random model

## Gaussian Kernel density estimation



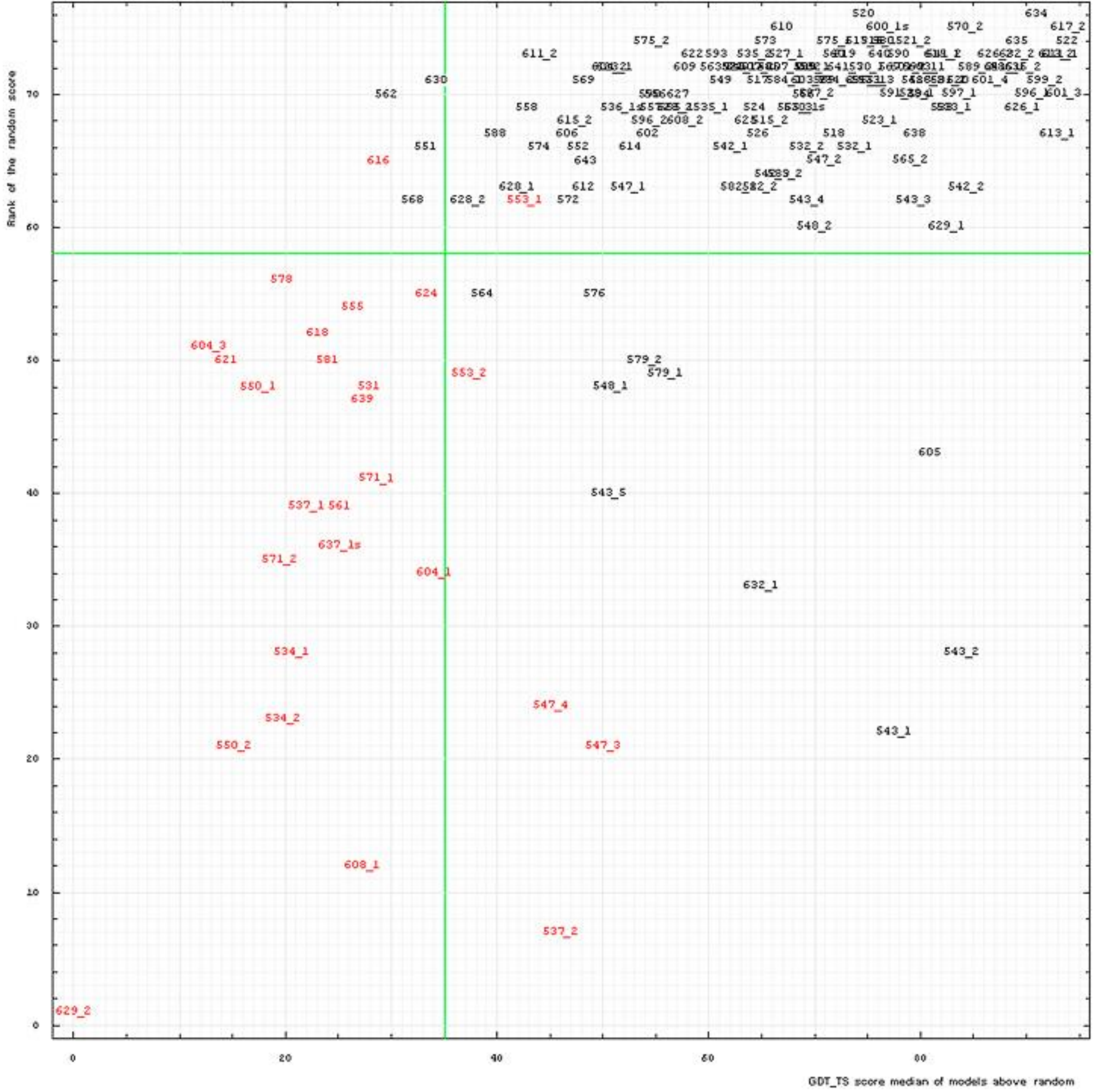
# 2D of these: CASP category definition

red: no HHpred templates



2D of these: CASP category definition

red: **FM targets**



# Acknowledgements

## Our group

**Lisa N. Kinch**
**ShuoYong Shi**
Jimin Pei
Qian Cong
Hua Cheng
Wenlin Li
Yuxing Liao
Dustin Schaeffer
Erik Nelson
Ming Tang
Jing Tong
Raquel Bromberg
Chalam Chitturi
Sasha Safronova
Bong-Hyun Kim
Jeremy Semeiks

**STRUCTURAL BIOLOGISTS for submitting CASP targets**

## CASP organizers

John Moult, CASP **president**, UM, USA
Krzysztof Fidelis, UC Davis, USA
Andriy Kryshtafovych, UC Davis, USA
Anna Tramontano, U of Rome, Italy

## CASP9 assessors:

Torsten Schwede, UBasel, Switzerland
Ken Dill, UCSF, USA
Justin MacCallum, UCSF, USA

## HHMI, NIH, UTSW, Welch Foundation