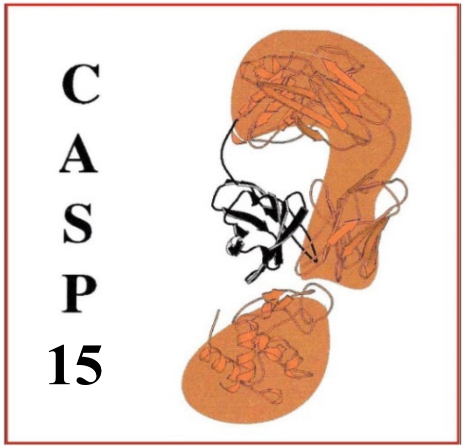


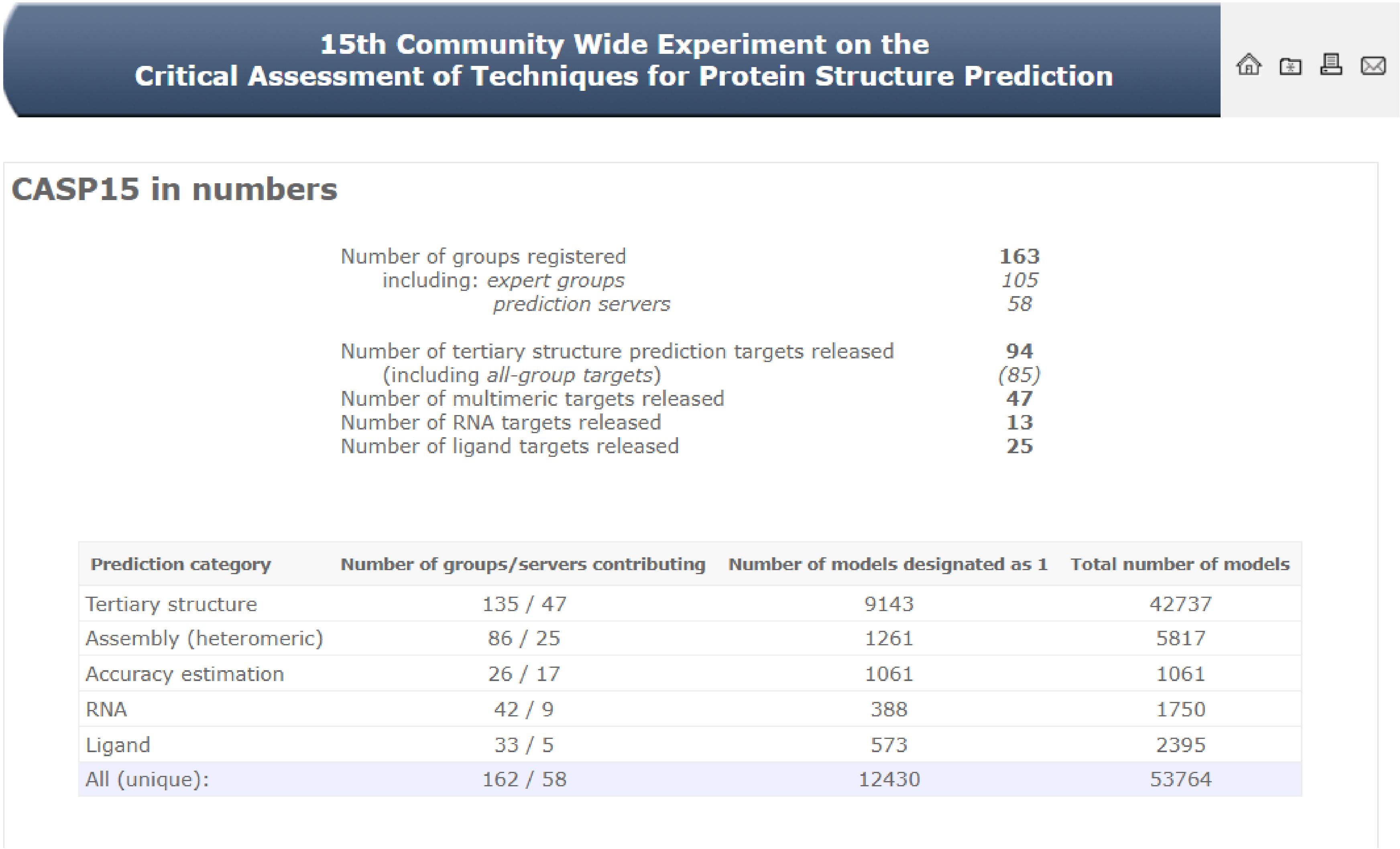
CASP15 in numbers

Andriy Kryshtafovych

Protein Structure Prediction Center
Genome Center
University of California, Davis
www.predictioncenter.org

Funded by the NIH/NIGMS

www.predictioncenter.org/casp15/numbers.cgi



15th Community Wide Experiment on the
Critical Assessment of Techniques for Protein Structure Prediction

CASP15 in numbers

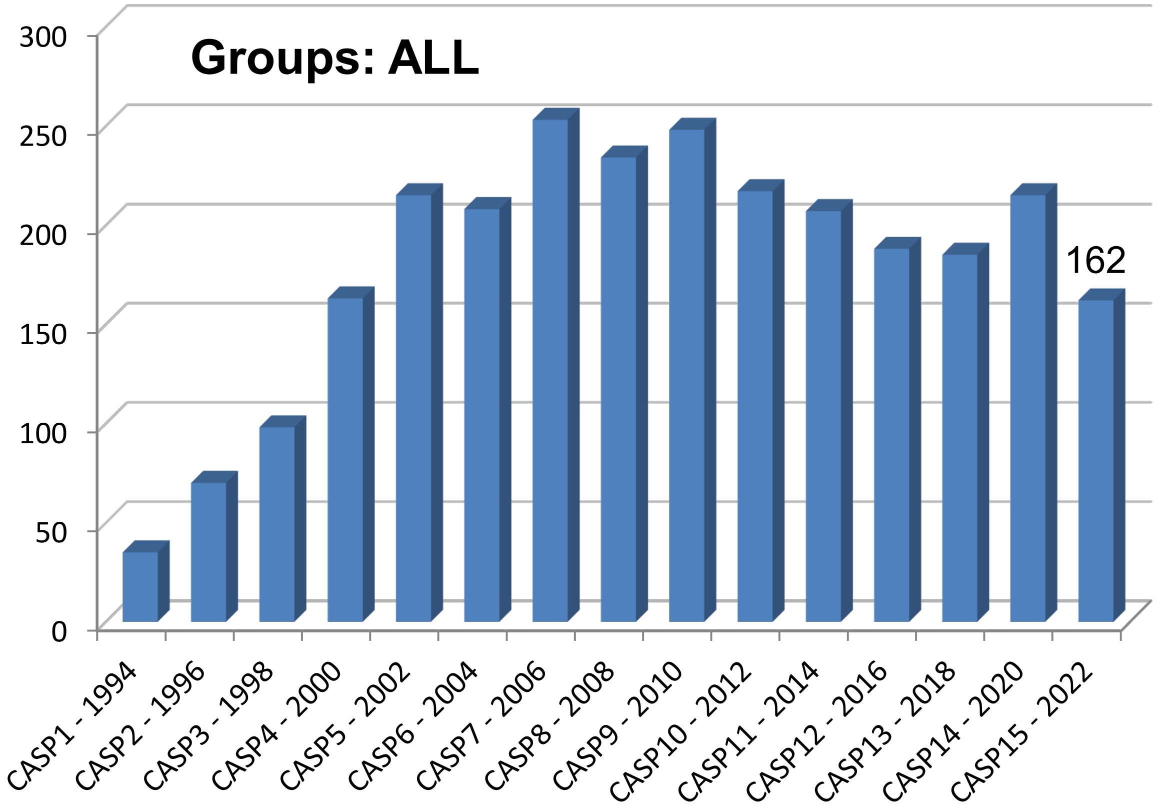
Number of groups registered	**163**
including: *expert groups*	*105*
prediction servers	*58*
Number of tertiary structure prediction targets released	**94**
(including *all-group targets*)	*(85)*
Number of multimeric targets released	**47**
Number of RNA targets released	**13**
Number of ligand targets released	**25**

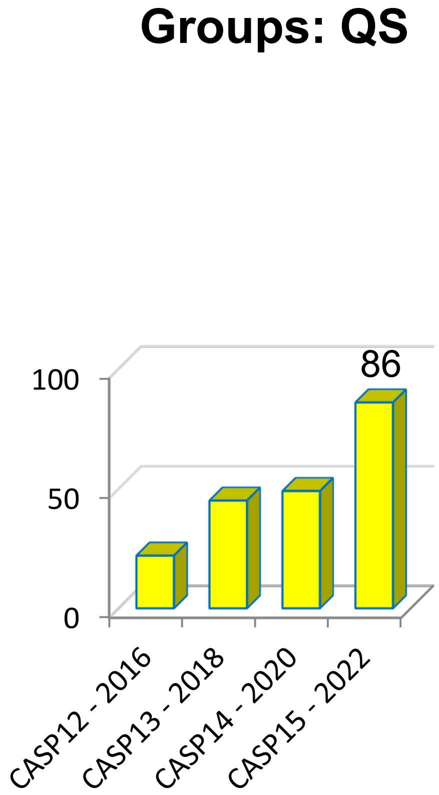
Prediction category	Number of groups/servers contributing	Number of models designated as 1	Total number of models
Tertiary structure	135 / 47	9143	42737
Assembly (heteromeric)	86 / 25	1261	5817
Accuracy estimation	26 / 17	1061	1061
RNA	42 / 9	388	1750
Ligand	33 / 5	573	2395
All (unique):	162 / 58	12430	53764

Groups (162 from 89 centers)

TS:135 QS:86

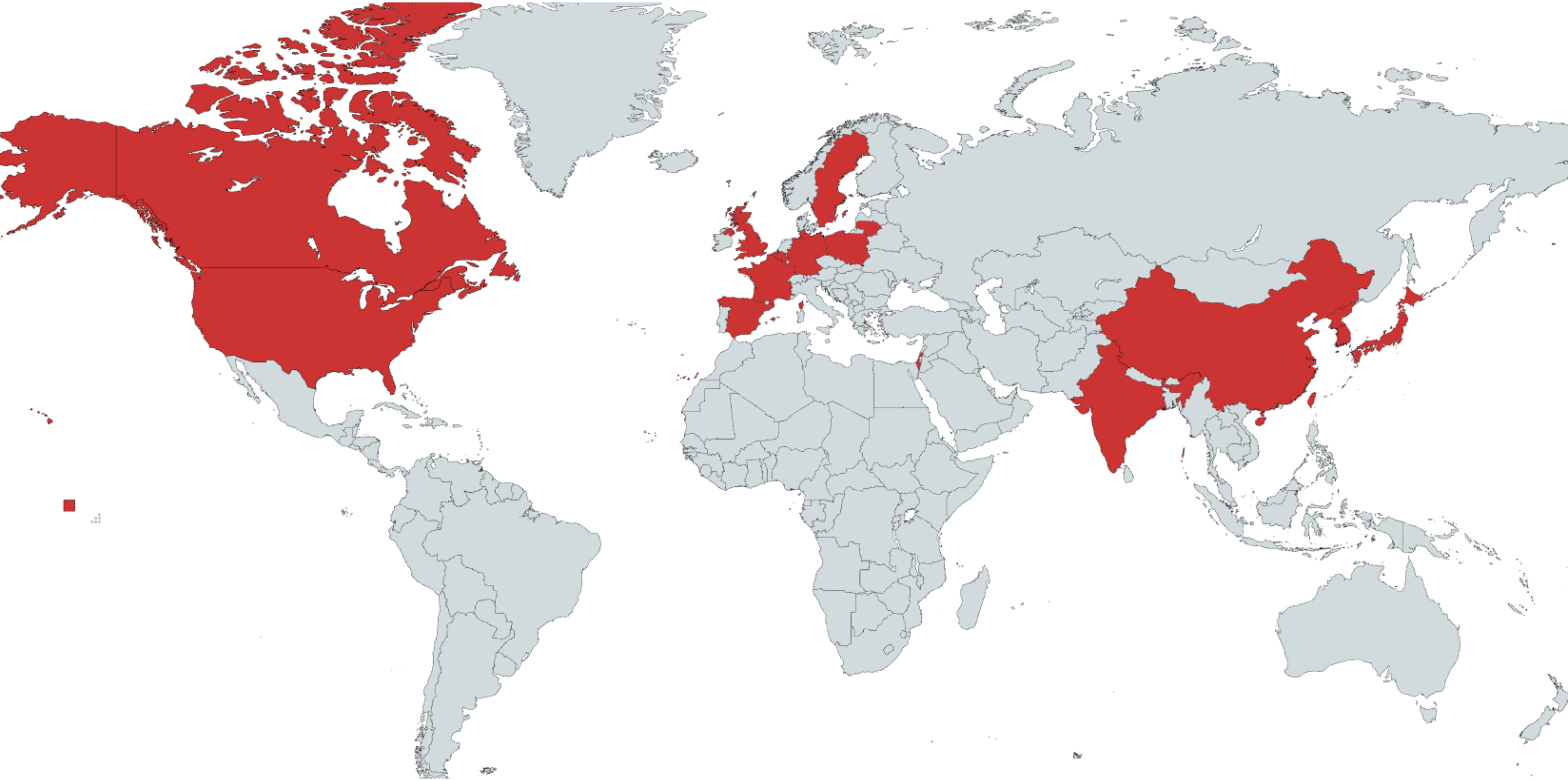
RNA:42 LIG:33 QA:26





CASP15 predictors geography

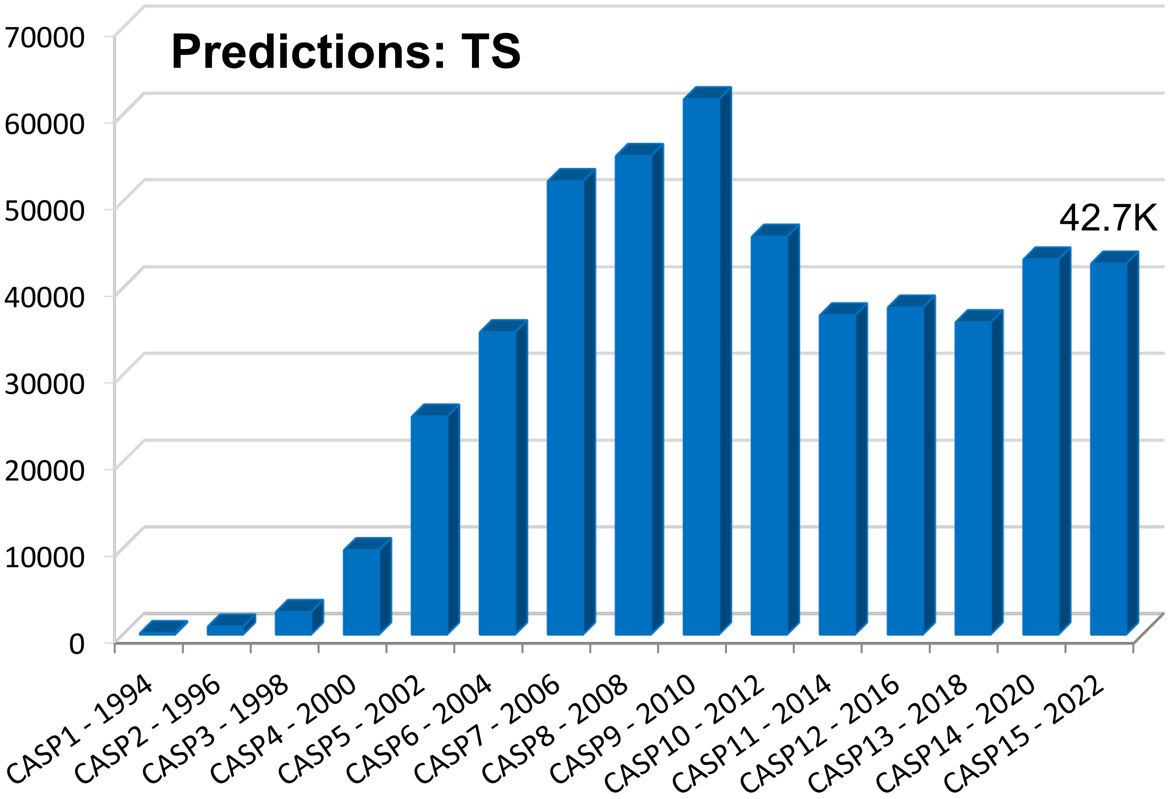
89 prediction centers from 17 countries
including
30 from the USA and 29 from China

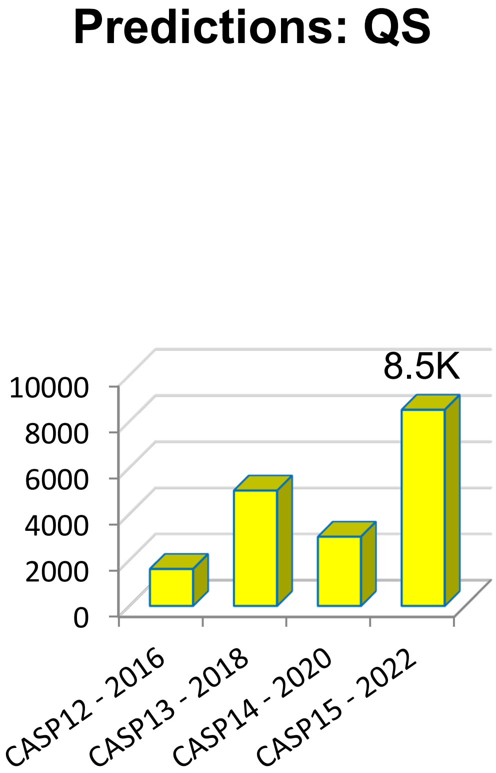


Predictions (53,700)

TS:42.7K QS:8.5K RNA:1.7K

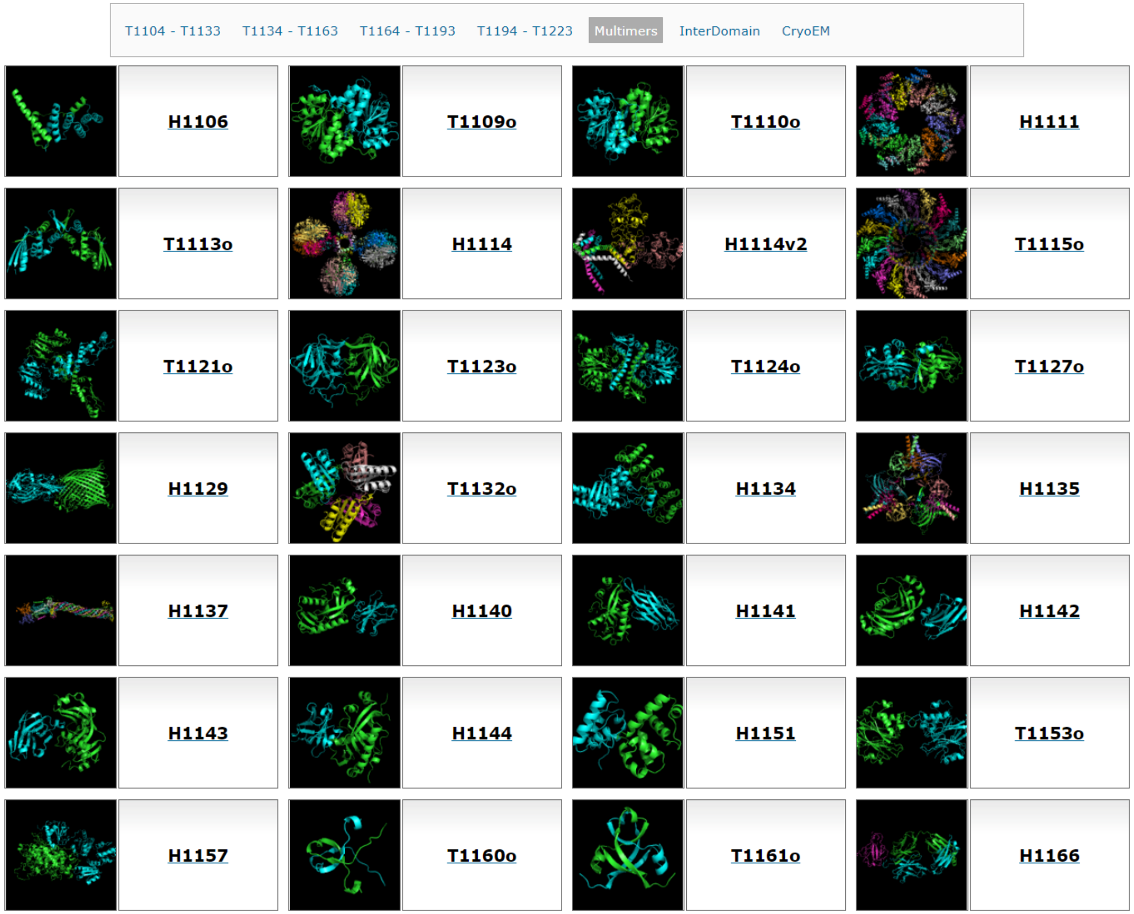
LIG:2.4K QA: 1.0K





Predictions (53,700)

(evaluation)



T1104 - T1133 T1134 - T1163 T1164 - T1193 T1194 - T1223 Multimers InterDomain CryoEM

H1106	T1109o	T1110o	H1111
T1113o	H1114	H1114v2	T1115o
T1121o	T1123o	T1124o	T1127o
H1129	T1132o	H1134	H1135
H1137	H1140	H1141	H1142
H1143	H1144	H1151	T1153o
H1157	T1160o	T1161o	H1166

Evaluation:

>30 different software tools including new ones:

SWORD (JC Gelly)

Foldseek (Martin Steinegger, Milot Mirdita)

reLLG (Randy Read)

USalign (Yang Zhang)

ASE-score*

QS-score*, including DockQ (Gabriel Studer)

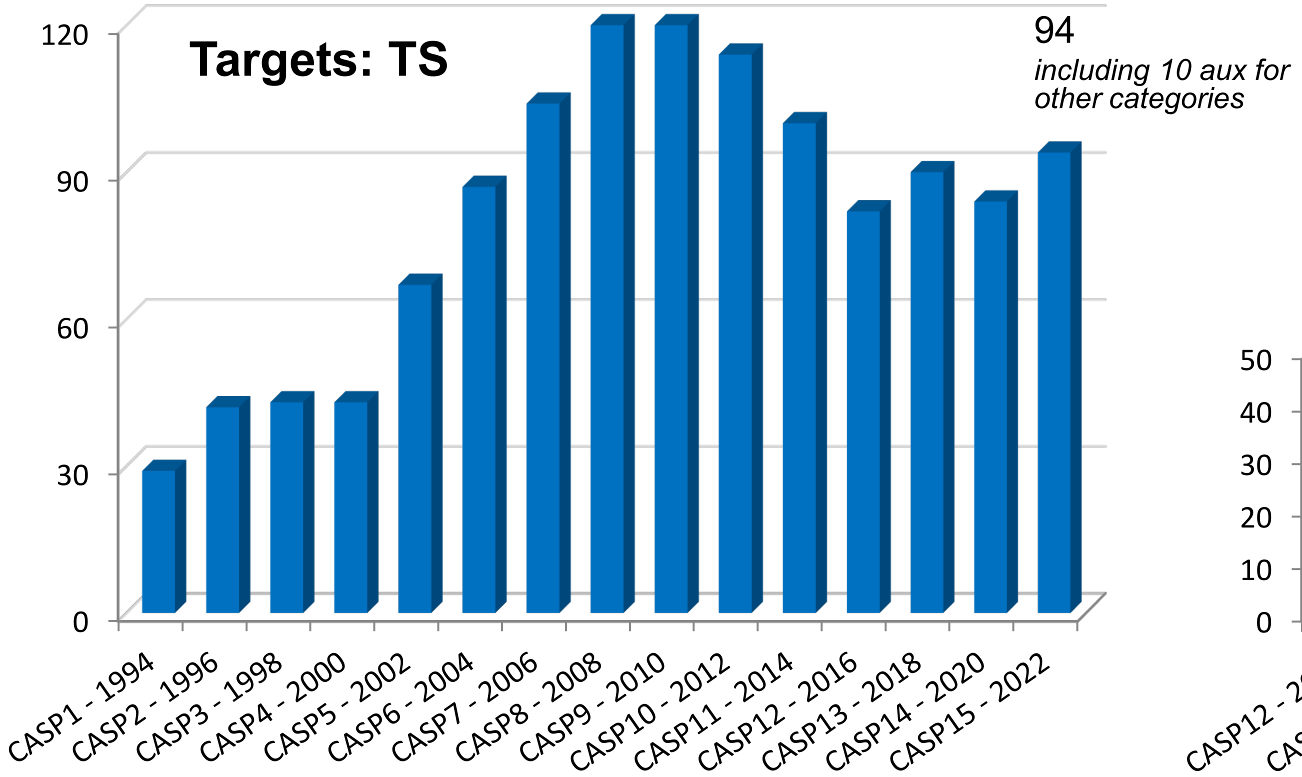
>20 visualization tools

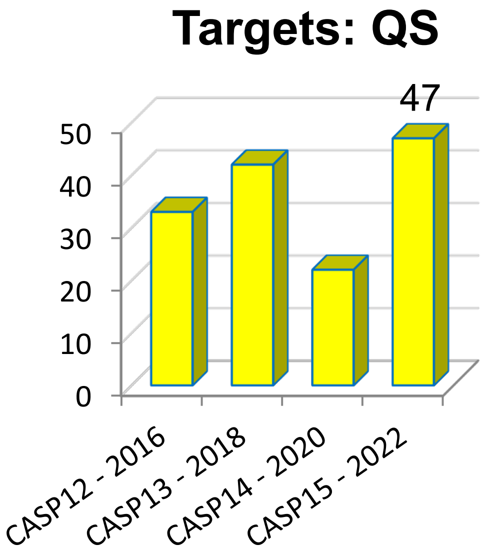
CASP15: 400 GB of data

Targets (96 entries)

TS:84 (77 assessed)   QS:47 (41 assessed, 43 assembly EUs)

RNA:12   LIG:23   ALT_CONF:14

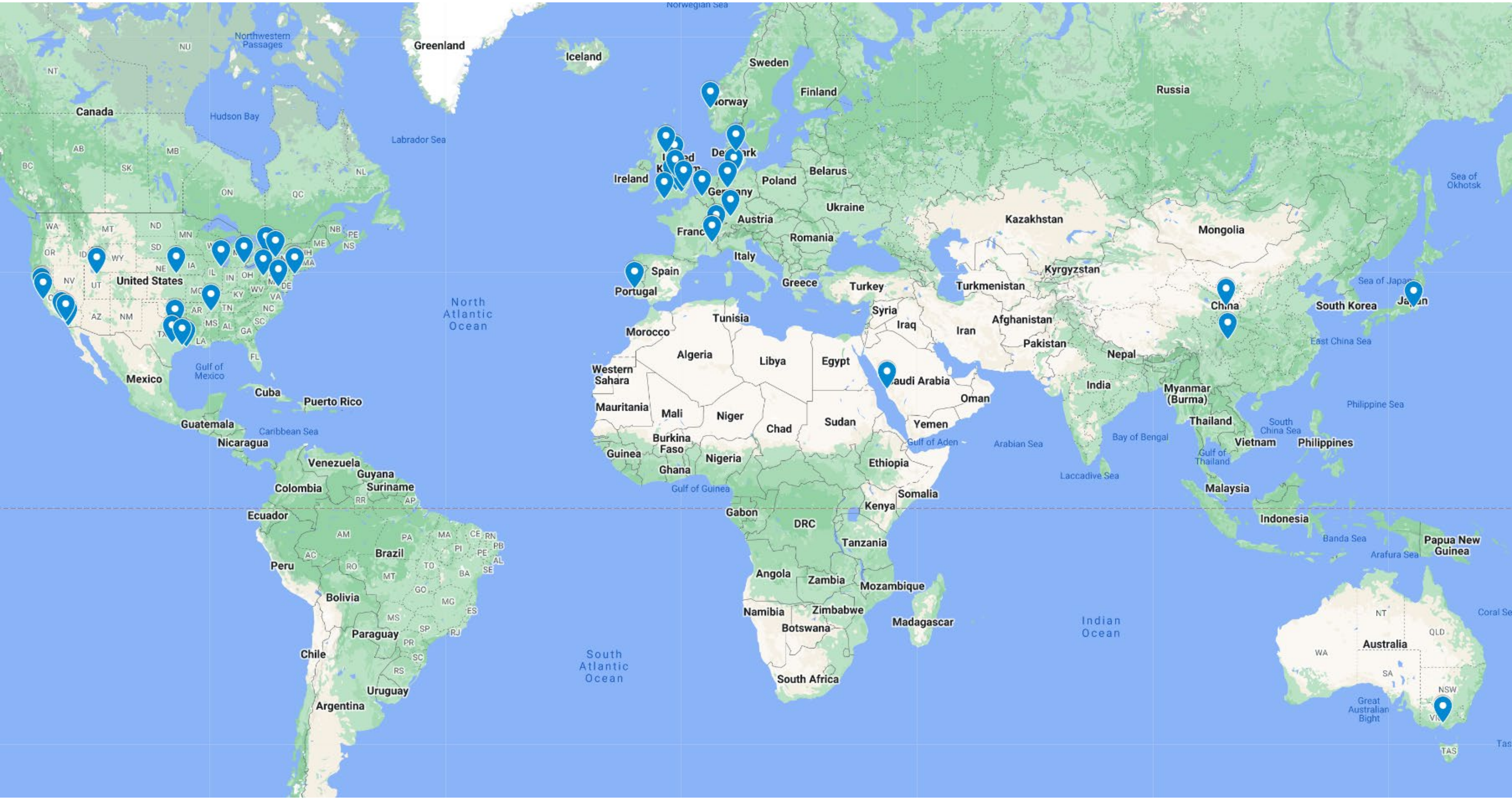




48 structure determination groups from 14 countries

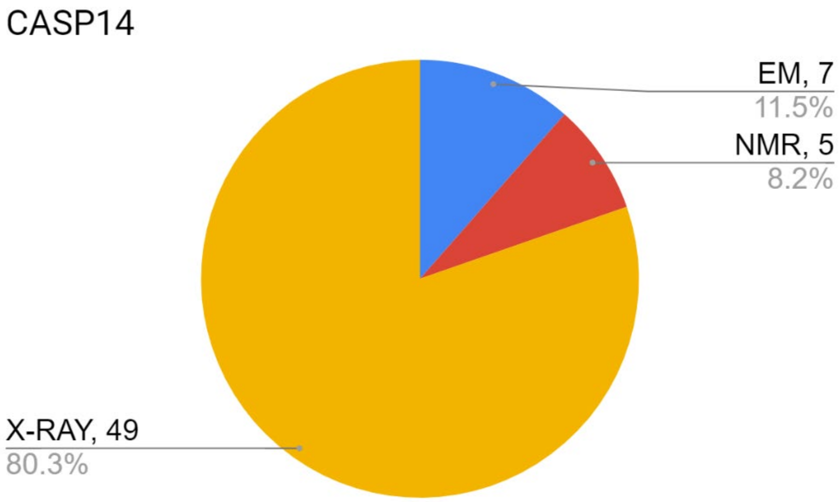
including

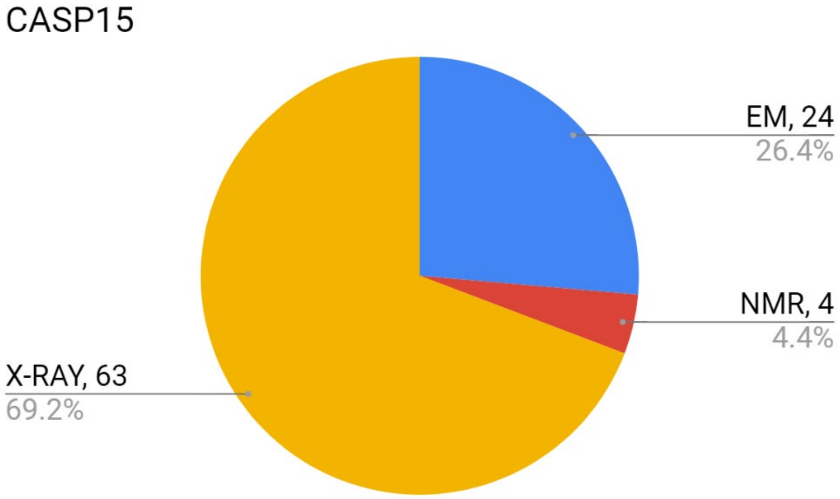
23 from the USA; 8 from the UK; 4 from Germany



Targets assessed (91 entries)

(determination method)

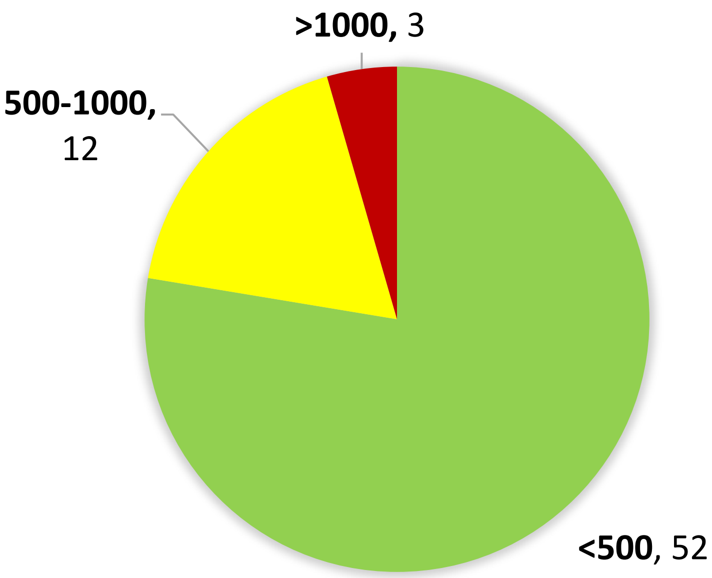




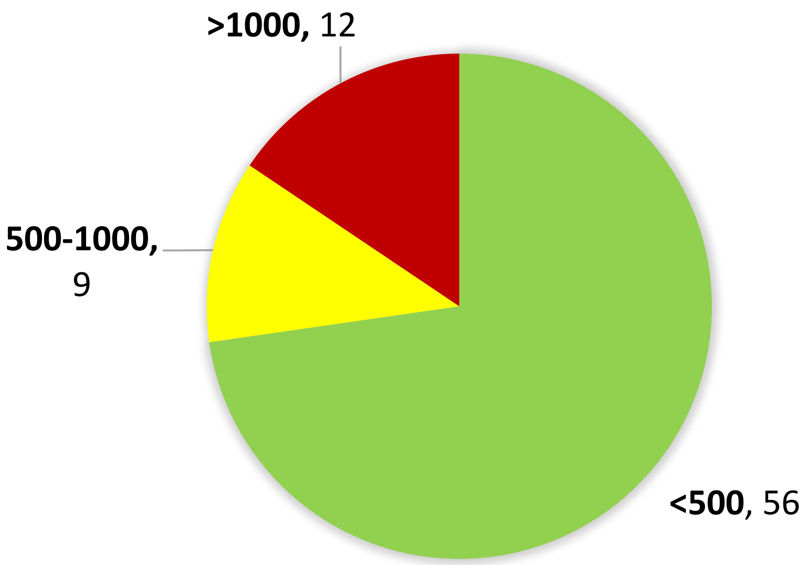
Protein targets assessed (77 entries)

(length)

CASP14



CASP15



Domain definition and classification

(together with Daniel Rigden)

1. Pre-processed targets as soon as structures become available.
2. Run domain boundary definition programs (DDomain, DomainParser2, SWORD).
3. Compare results of homology search programs (PSIBLAST, HHsearch) with #2.
4. Suggested preliminary domain definition based on #2, #3 and visual inspection.
5. Run evaluation of models and template search for the suggested domains.
6. Suggested composition of evaluation units (EUs) based on the domain-based evaluation results (Grishin plots) and, if needed,

   6.1. - rerun evaluation on the adjusted EUs (minimal number of exceptions).
7. Classified domains in 4 difficulty categories, TBM-easy, TBM-hard, TBM/FM, FM based on the homology searches. *(This was different from CASP14 where this classification was done based on the extensive manual examination of homology relations (Lisa Kinch) and performance of modelers).*

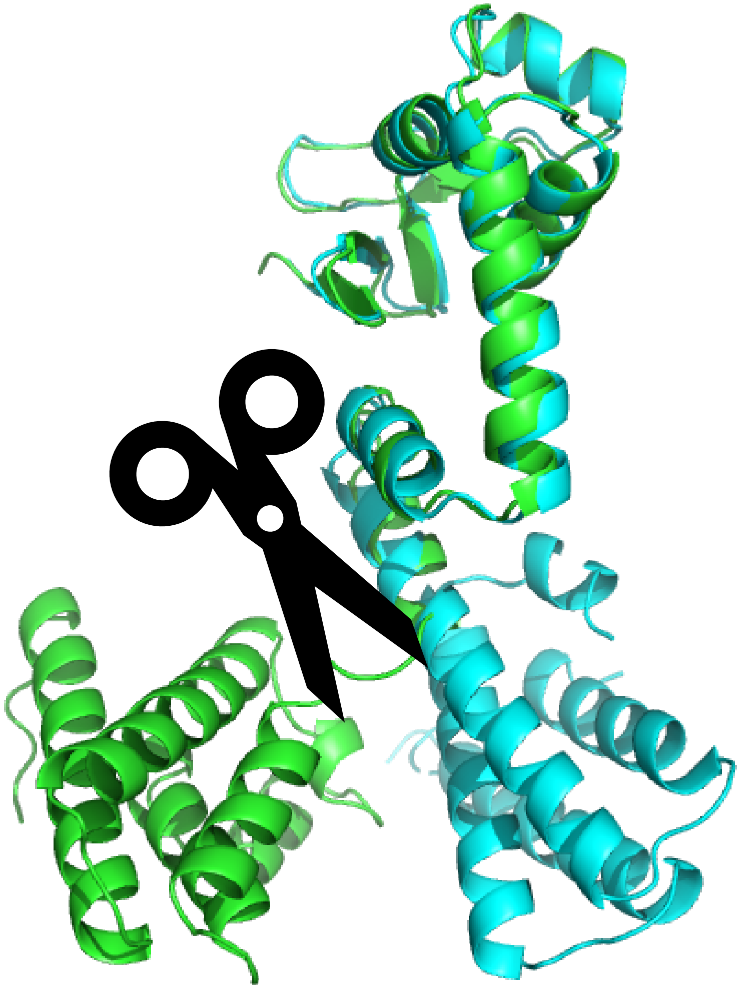
Domain definition

(have to split)

T1120

DNA-binding protein:

a monomer with two chains in the ASU and different orientation of domains



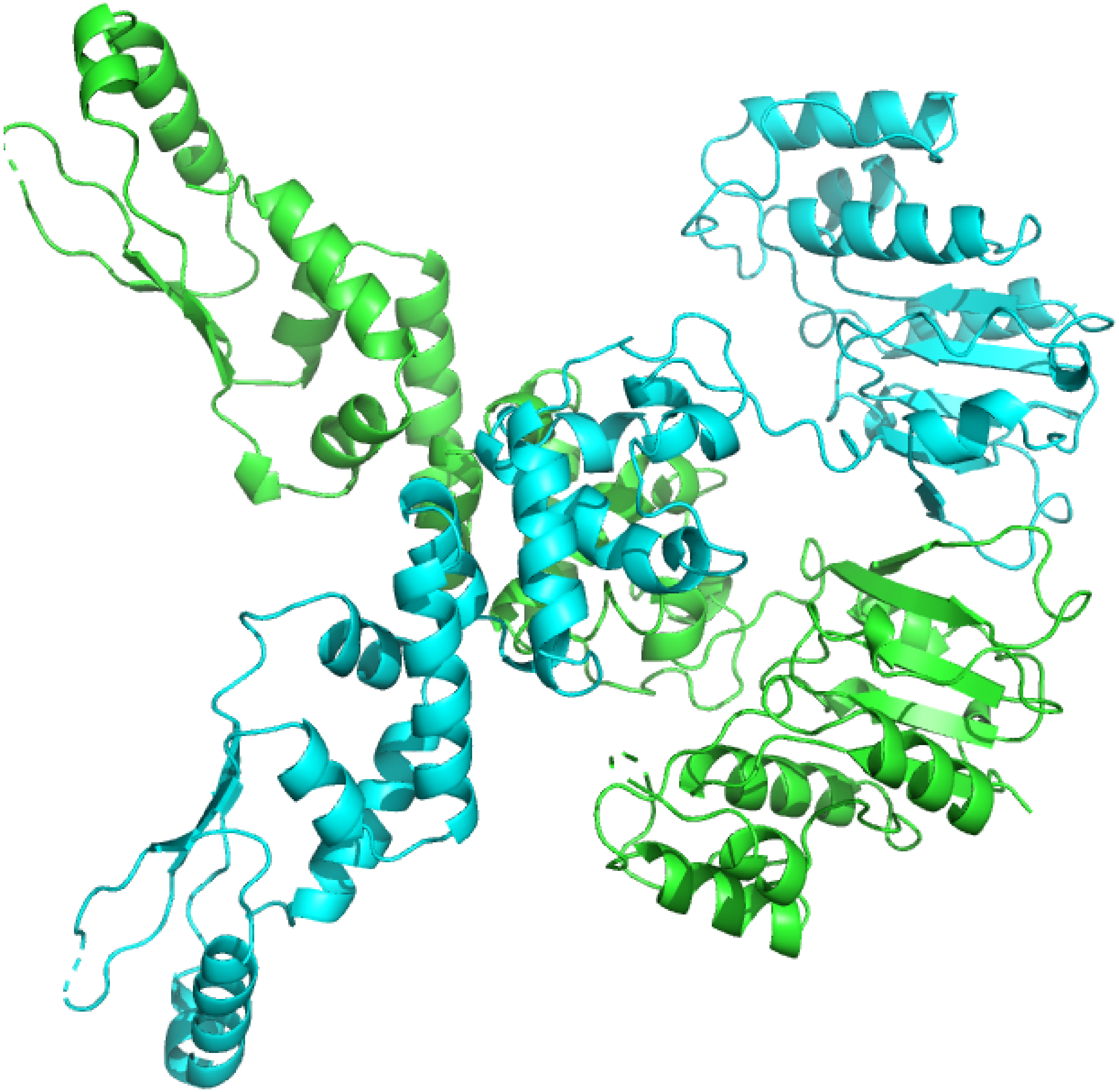
Domain definition

(have to split)

T1121

DNA cleavage protein:

a homodimer with different orientation of domains in two chains



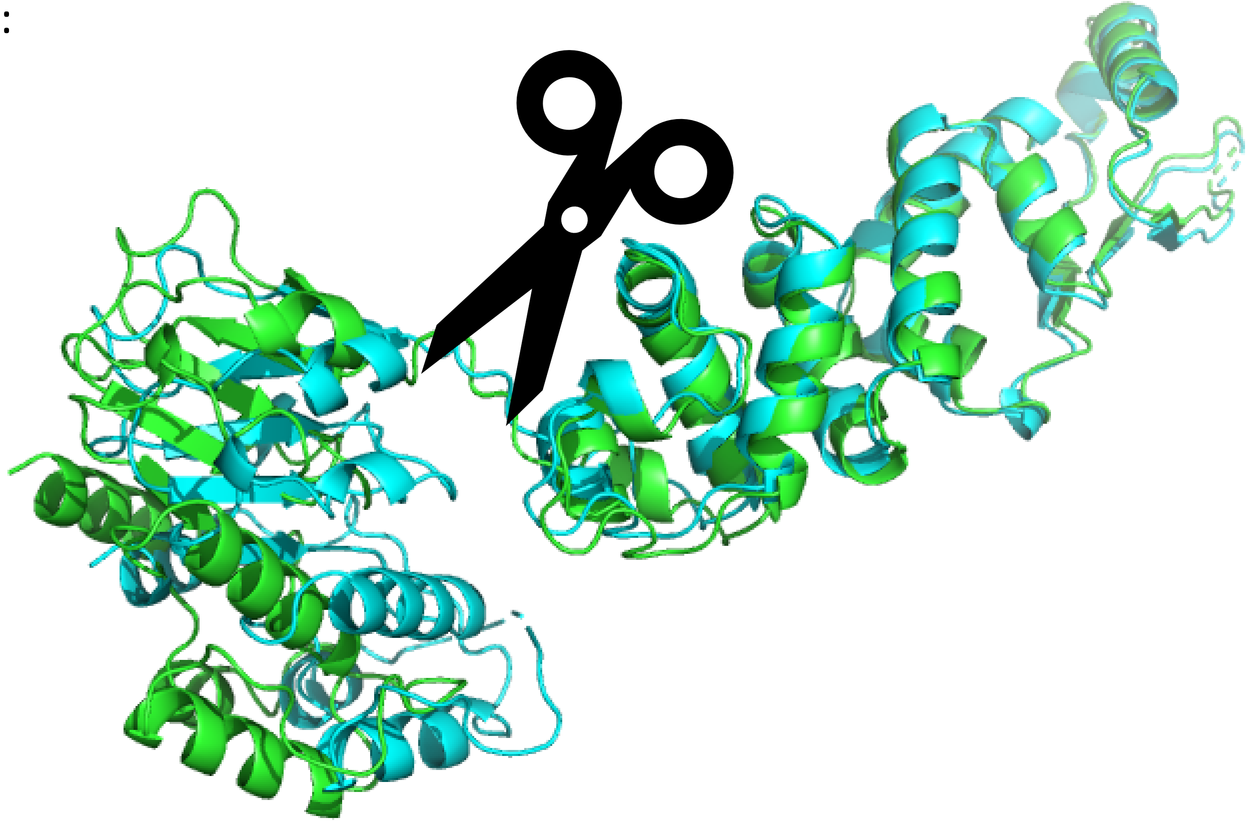
Domain definition

(have to split)

T1121

DNA cleavage protein:

a homodimer with different orientation of domains in two chains



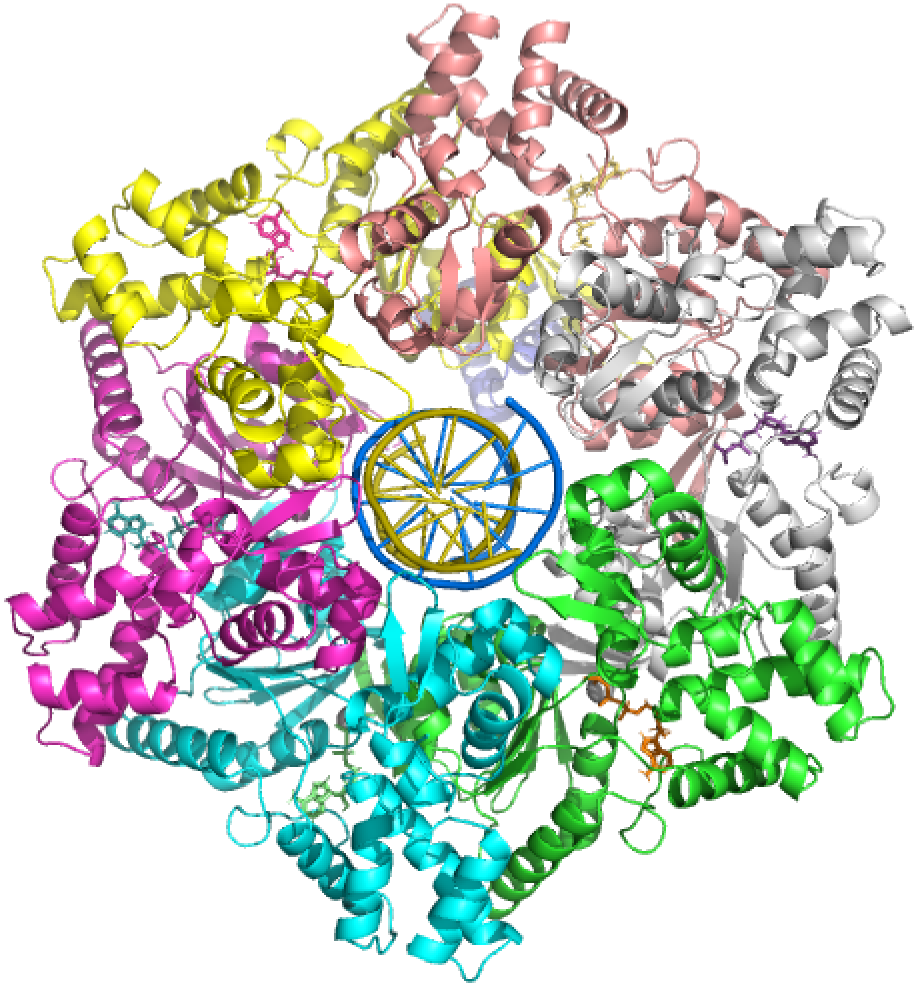
Domain definition

(have to split)

T1170

the Holiday junction hexamer with non-crystallographic symmetry:

some chains deform to accommodate DNA



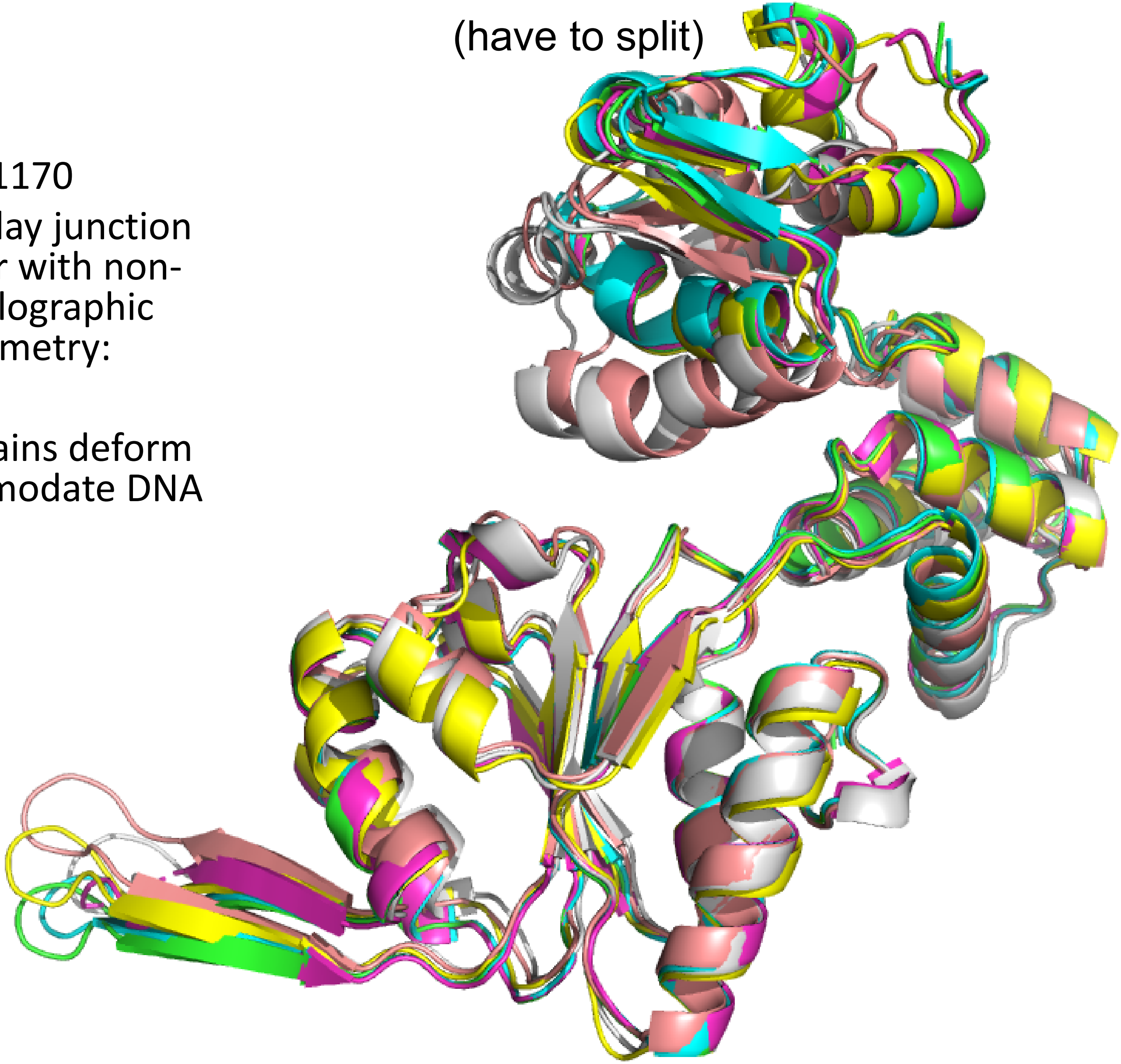
Domain definition

(have to split)

T1170

the Holiday junction hexamer with non-crystallographic symmetry:

some chains deform to accommodate DNA



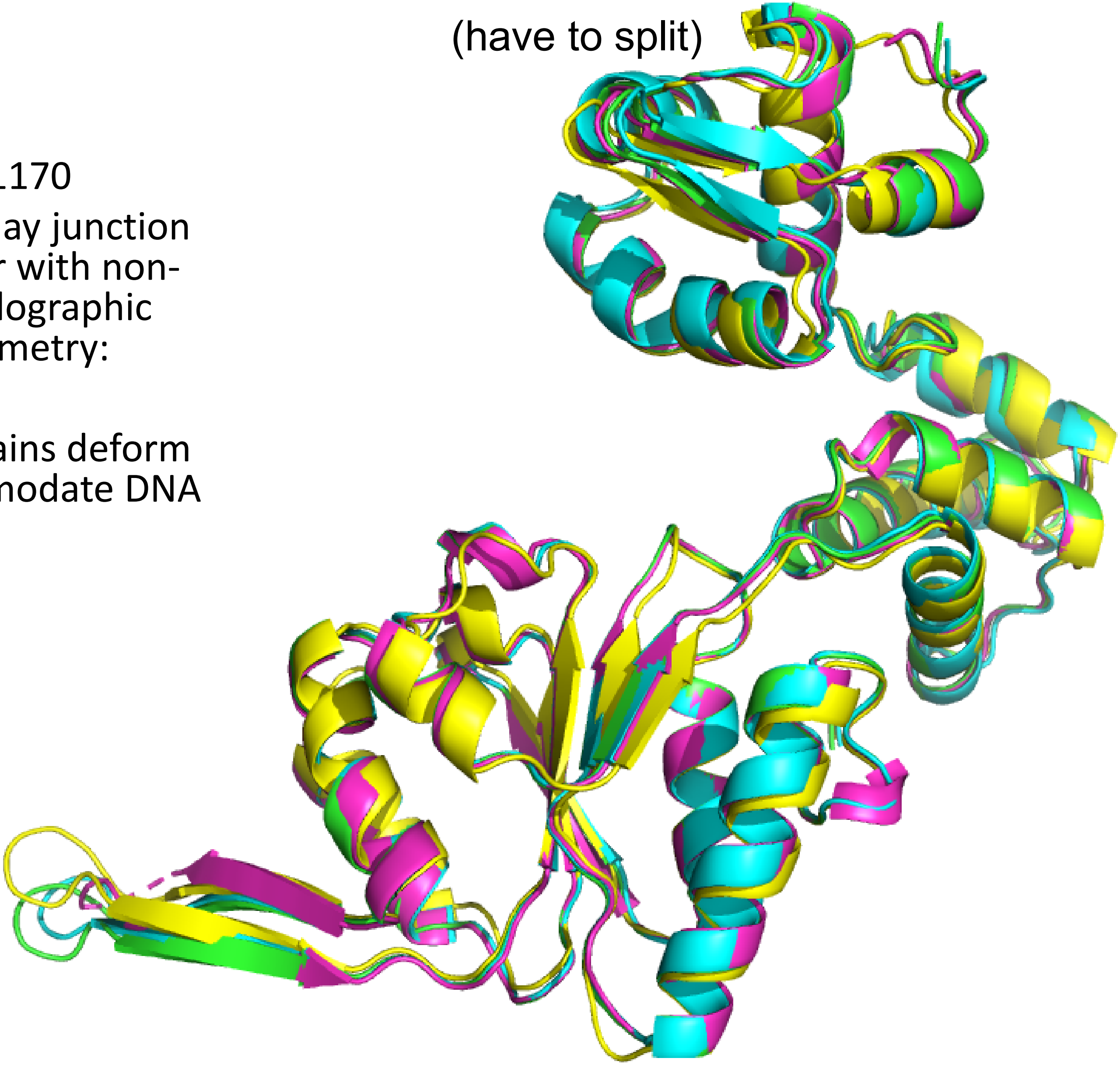
Domain definition

(have to split)

T1170

the Holiday junction hexamer with non-crystallographic symmetry:

some chains deform to accommodate DNA



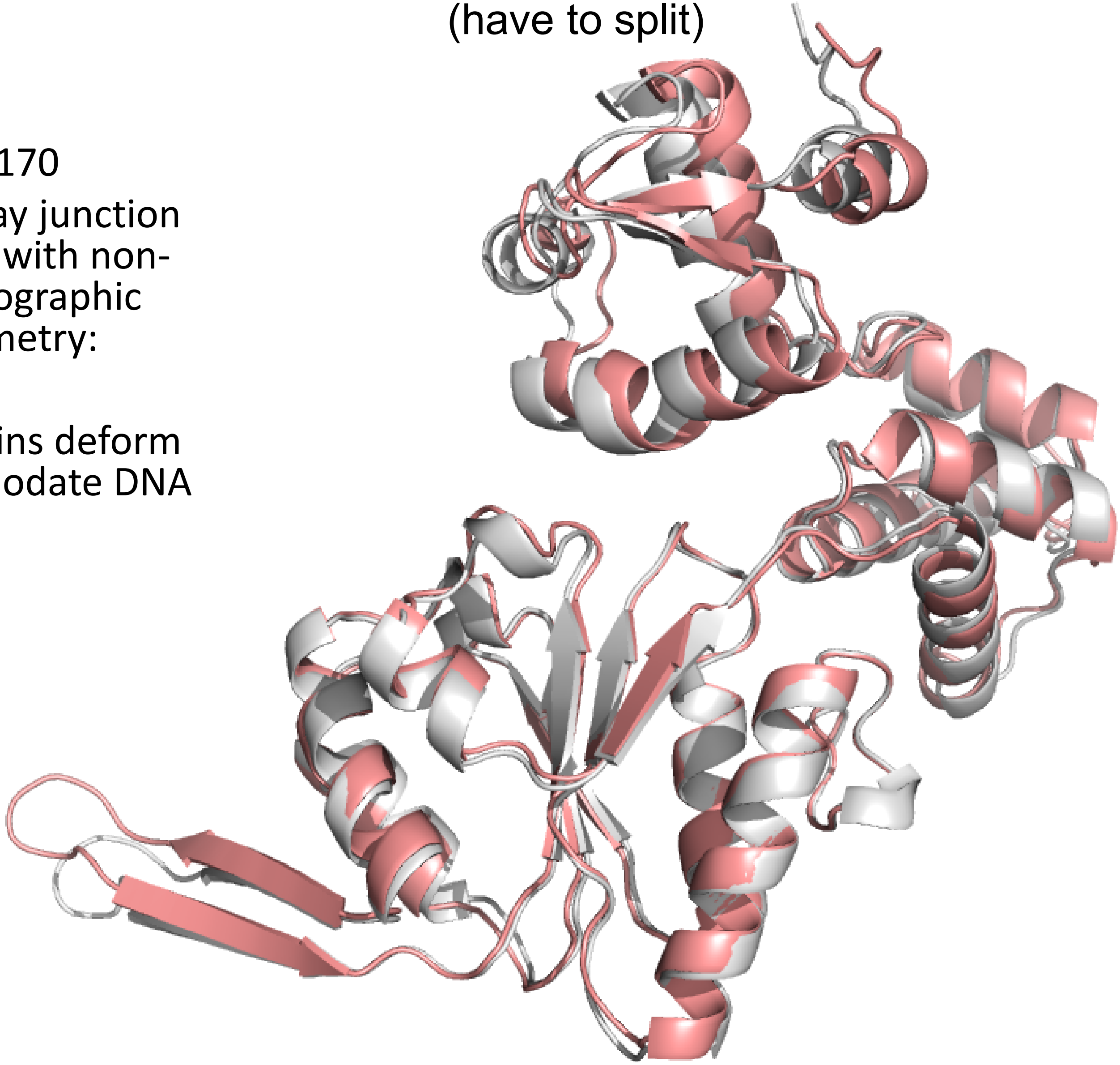
Domain definition

(have to split)

T1170

the Holiday junction hexamer with non-crystallographic symmetry:

some chains deform to accommodate DNA



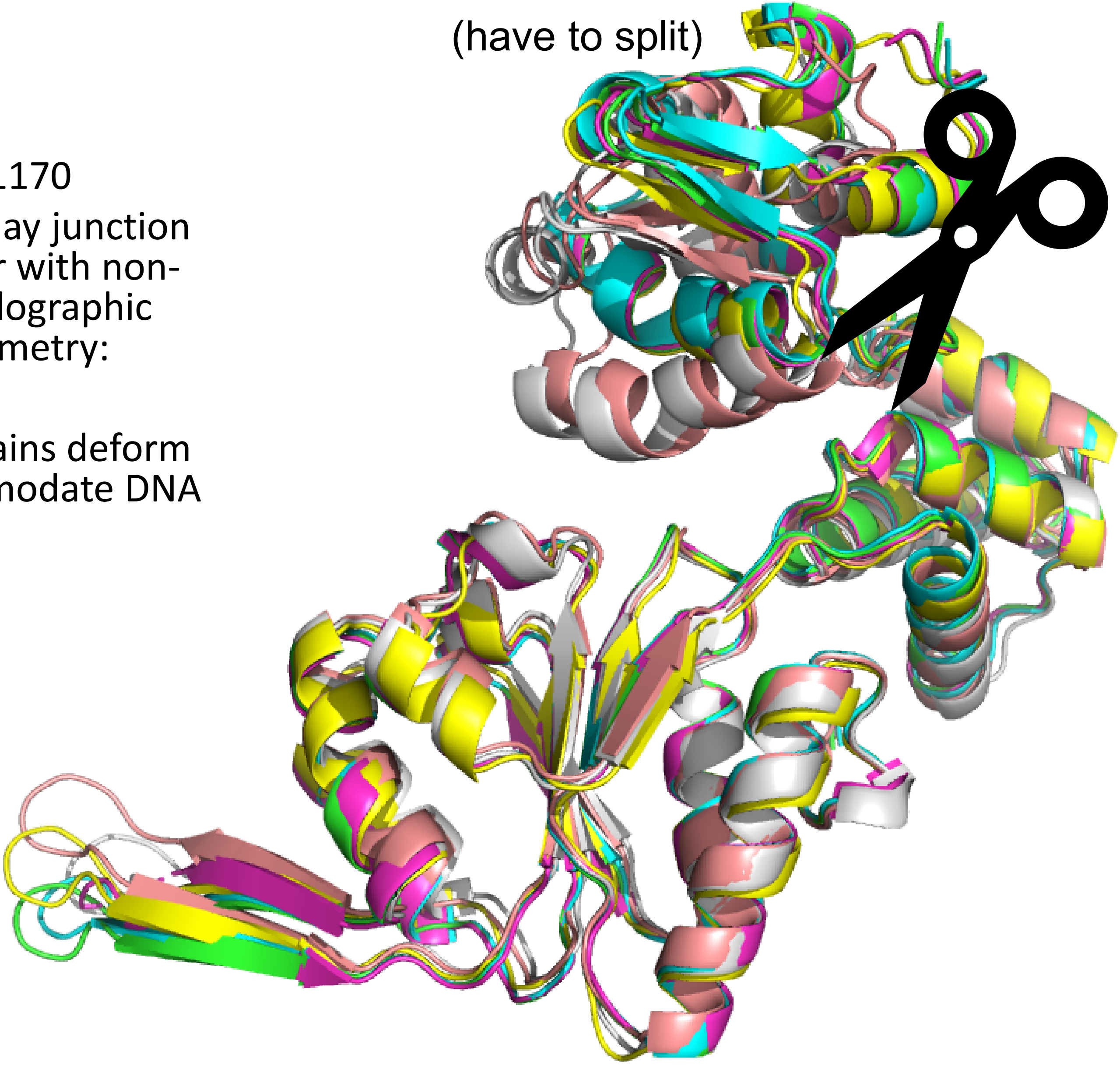
Domain definition

(have to split)

T1170

the Holiday junction hexamer with non-crystallographic symmetry:

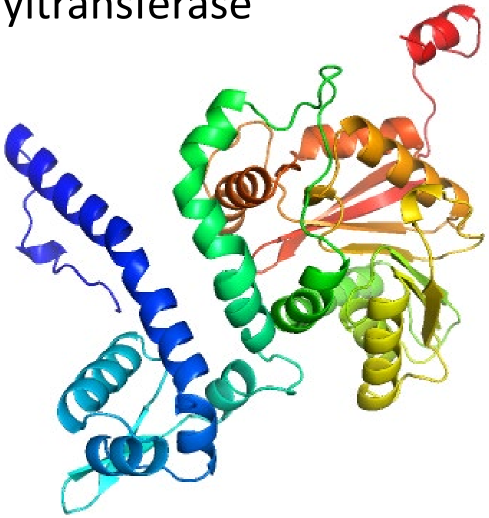
some chains deform to accommodate DNA



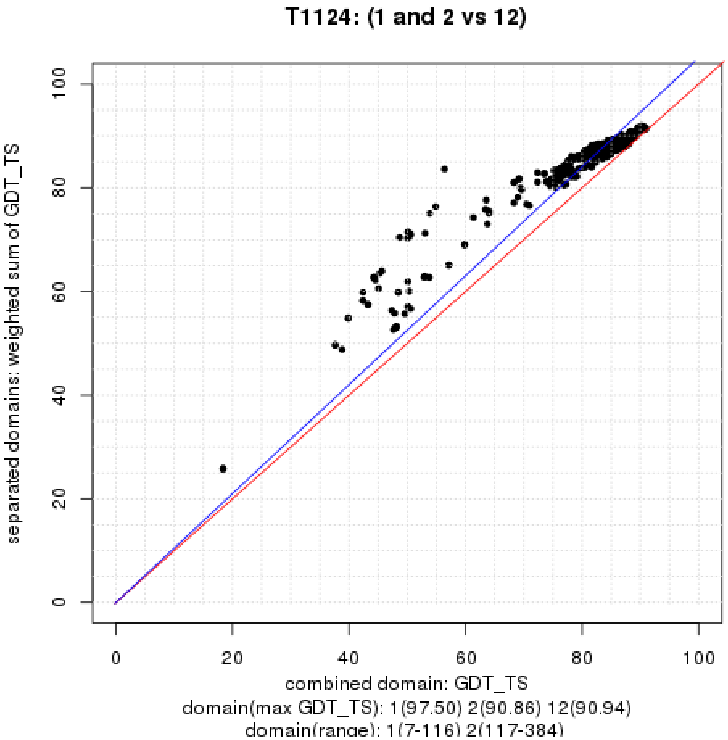
Domain definition

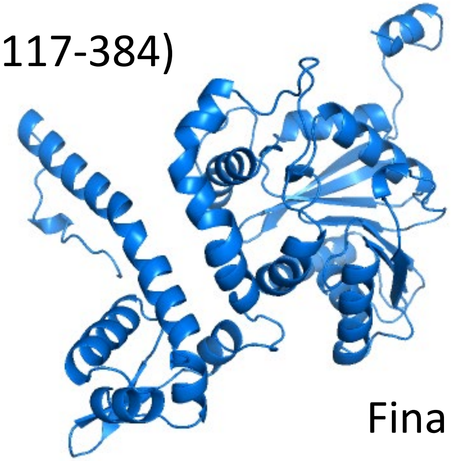
(to split or not to split)

T1124
methyltransferase



DDomain: 2 (7-116)(117-384)





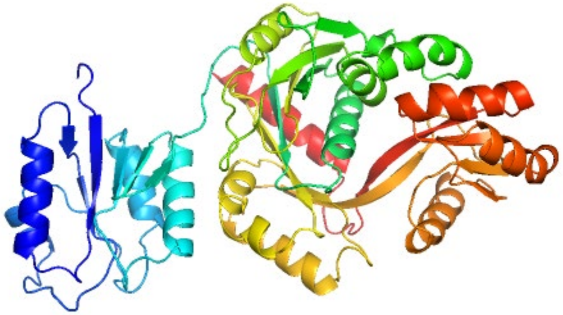
Final EU definition: 1 (7-384)

Domain definition

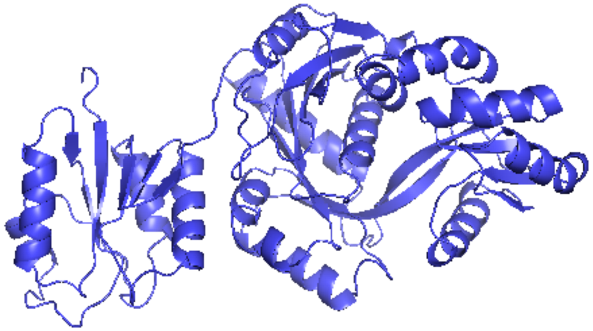
(to split or not to split)

T1112

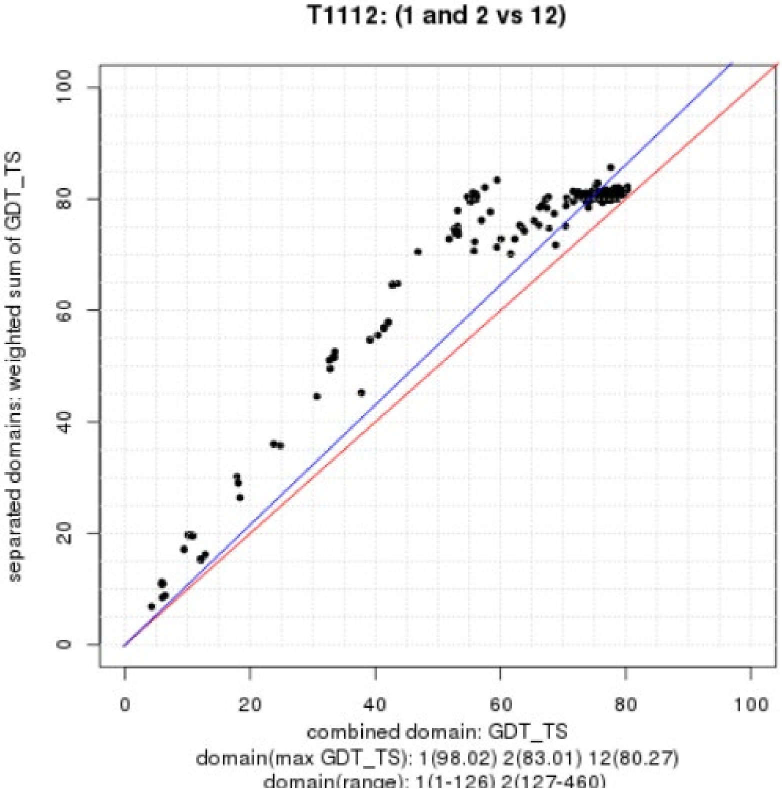
protein involved in the
synthesis of a rare osmolyte



DomainParser: 2 (1-126) (127-460)



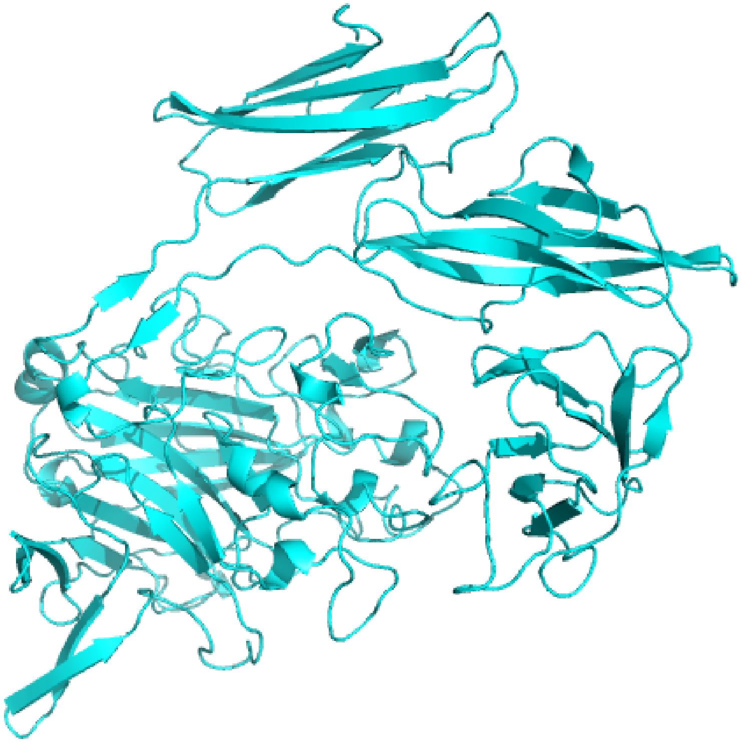
Final EU definition: 1   (1-460)

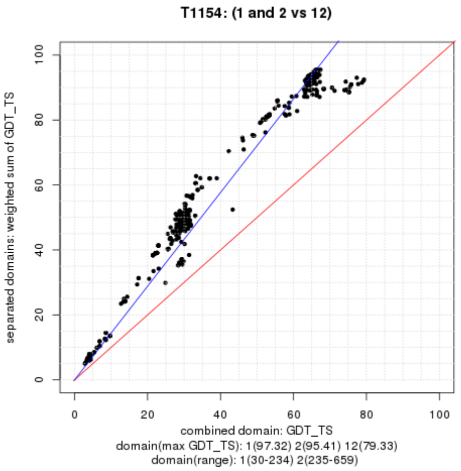


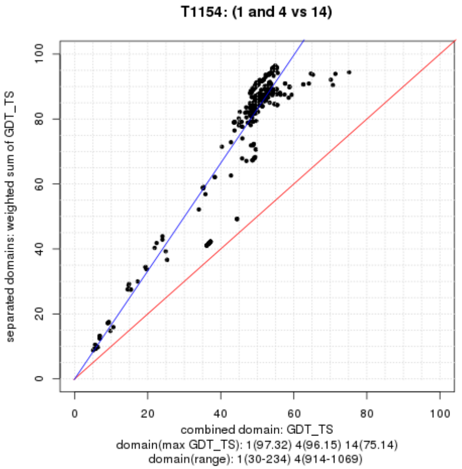
Domain definition

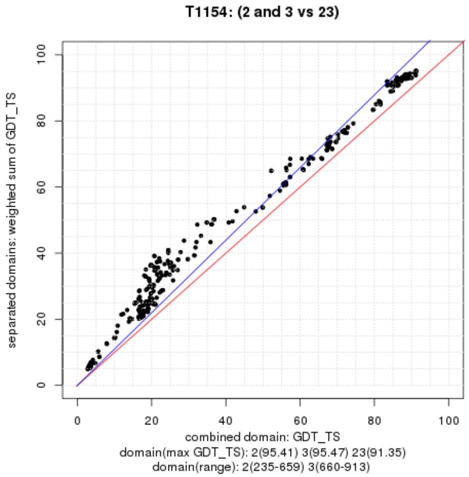
(to split or not to split)

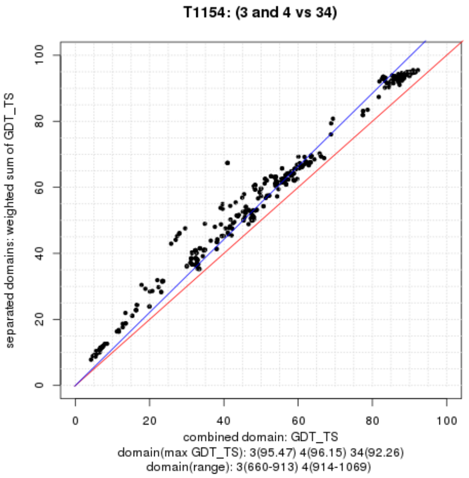
T1154
S-layer protein




T1154: (1 and 2 vs 12)
separated domains: weighted sum of GDT_TS
0
20
40
60
80
100
combined domain: GDT_TS
domain(max GDT_TS): 1(97.32) 2(95.41) 12(79.33)
domain(range): 1(30-234) 2(235-659)





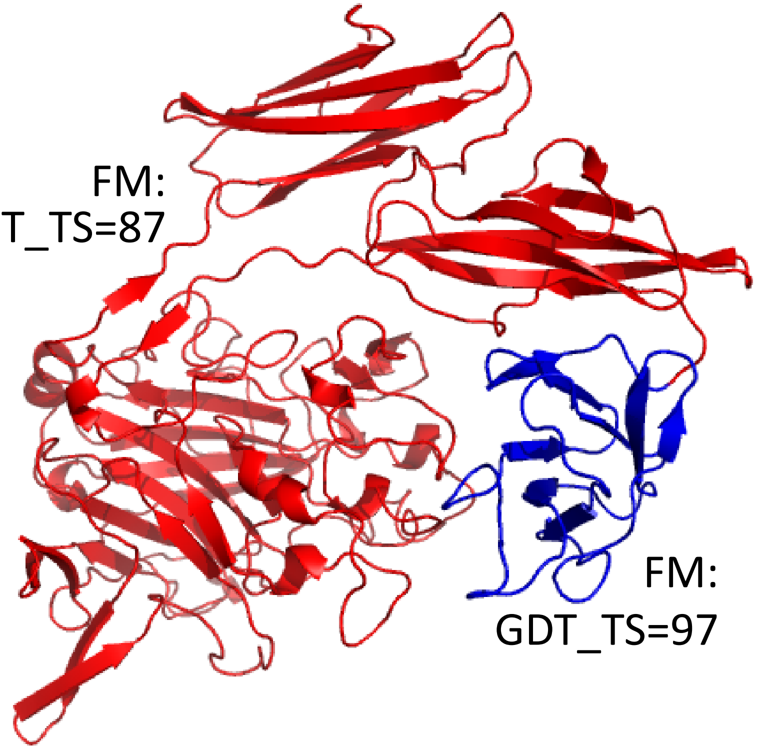


DDomain: 4 (30-234) (235-659) (660-913) (914-1069)

Domain definition

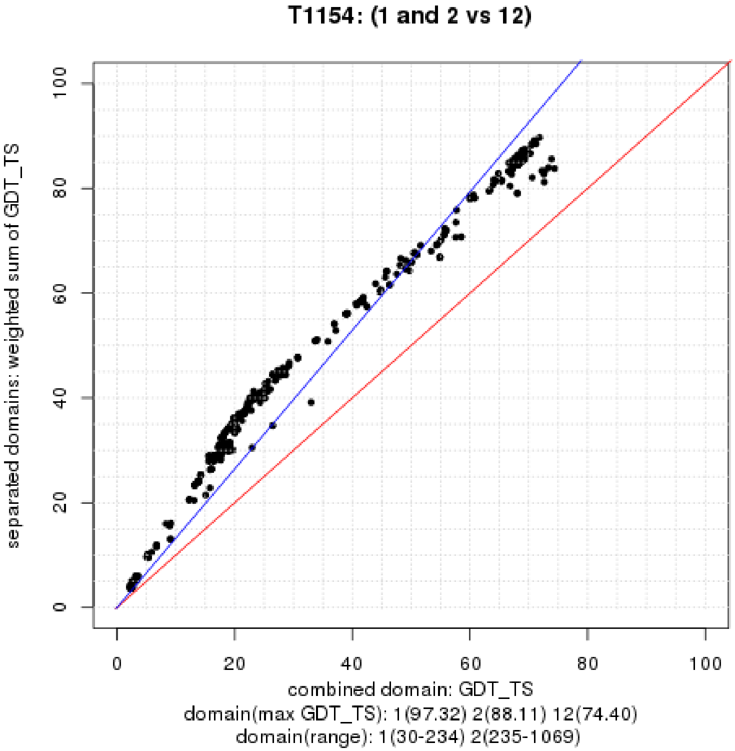
(to split or not to split)

T1154
S-layer protein



FM:
GDT_TS=87

FM:
GDT_TS=97



Started from 4 domains – ended up with 2EU: 2 (30-234)(235-1069)

Domain definition

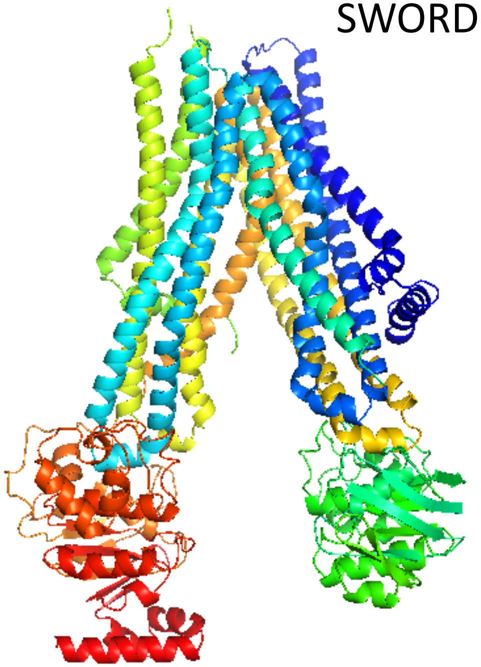
(to split or not to split)

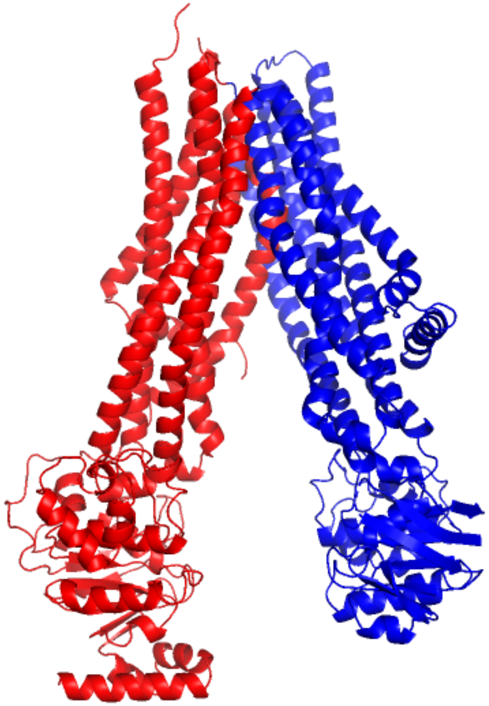
T1158

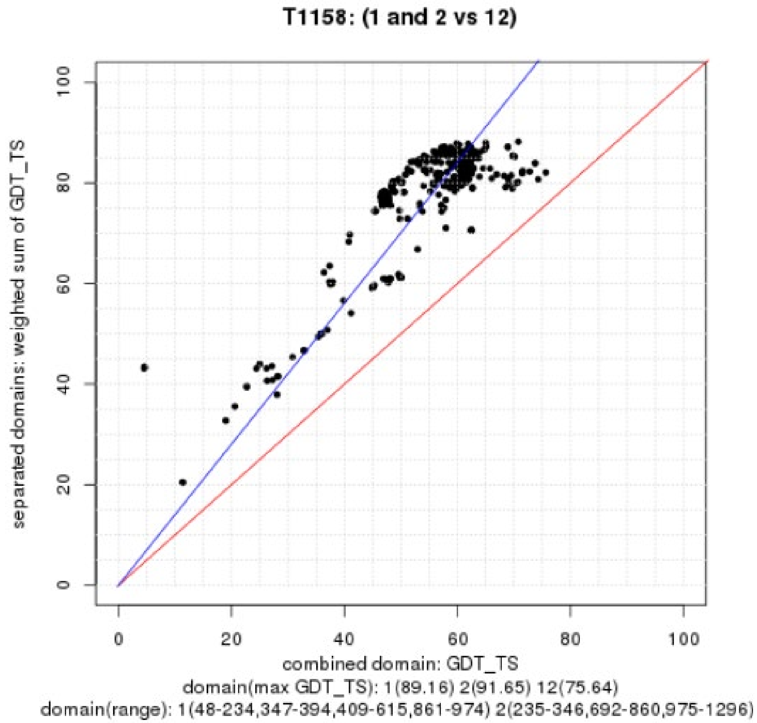
DomainParser: 2 (48-1022) (1023-1296)

DDomain: 6 (48-173)(174-394)(409-615)(692-796)(797-1021)(1022-1296)

SWORD: 5






T1158: (1 and 2 vs 12)
separated domains: weighted sum of GDT_TS
0
20
40
60
80
100
0
20
40
60
80
100
combined domain: GDT_TS
domain(max GDT_TS): 1(89.16) 2(91.65) 12(75.64)
domain(range): 1(48-234,347-394,409-615,861-974) 2(235-346,692-860,975-1296)

D1: 48-234,347-394,409-615,861-974

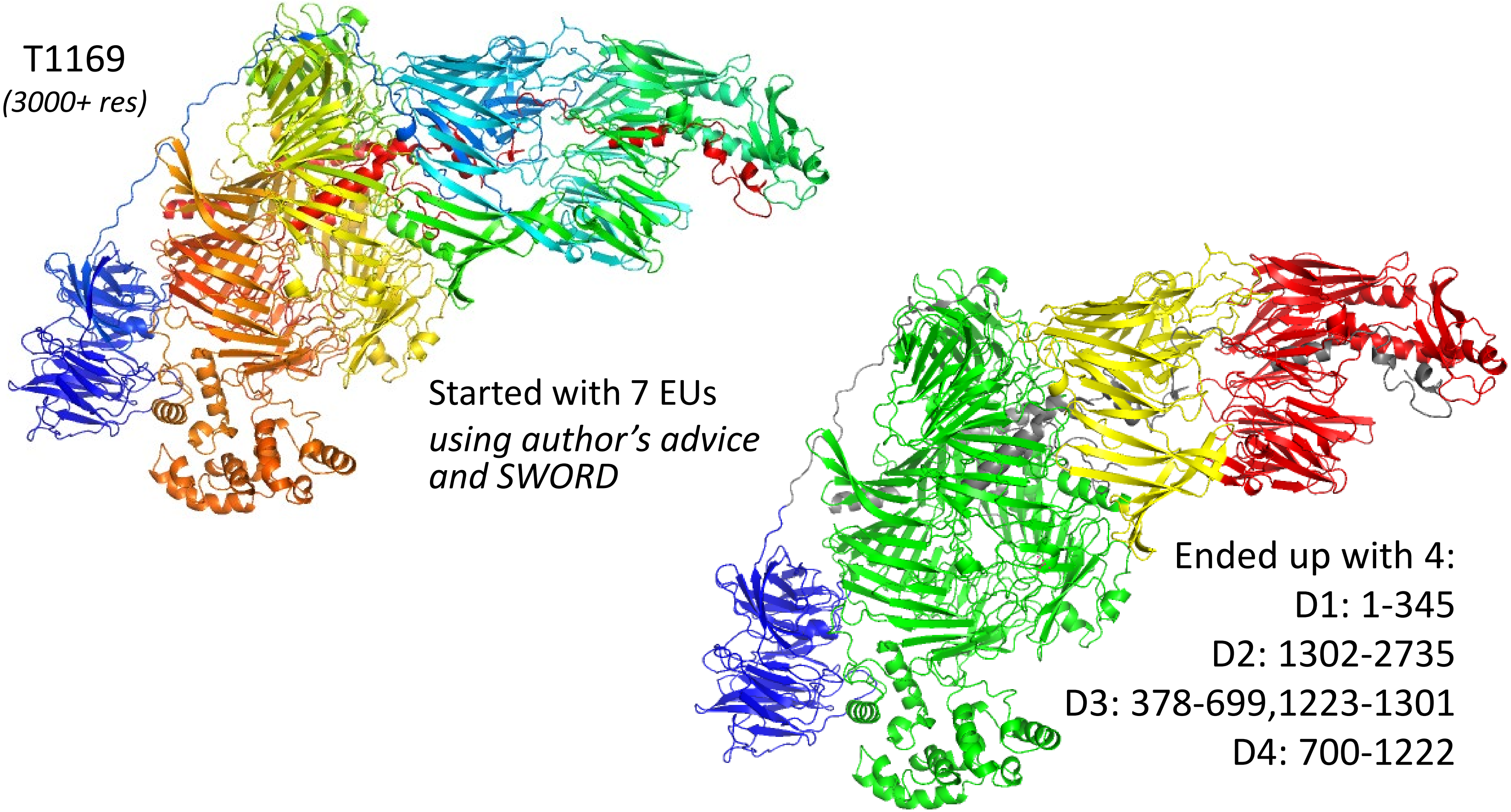
D2: 235-346,692-860,975-1296

Domain definition

(to split or not to split)

T1169

(3000+ res)



Started with 7 EUs

using author's advice and SWORD

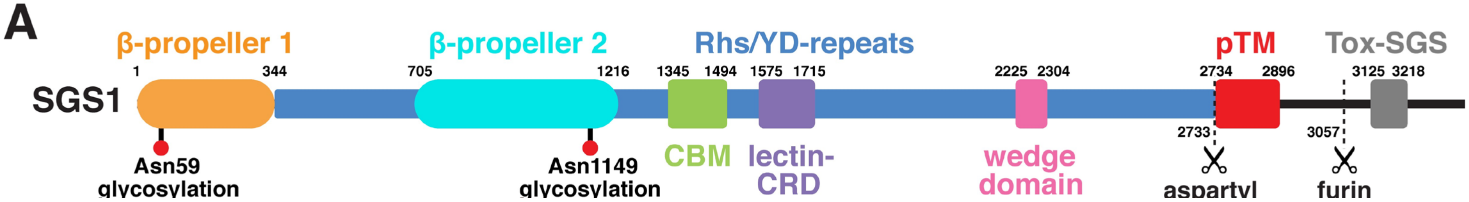
Ended up with 4:

D1: 1-345

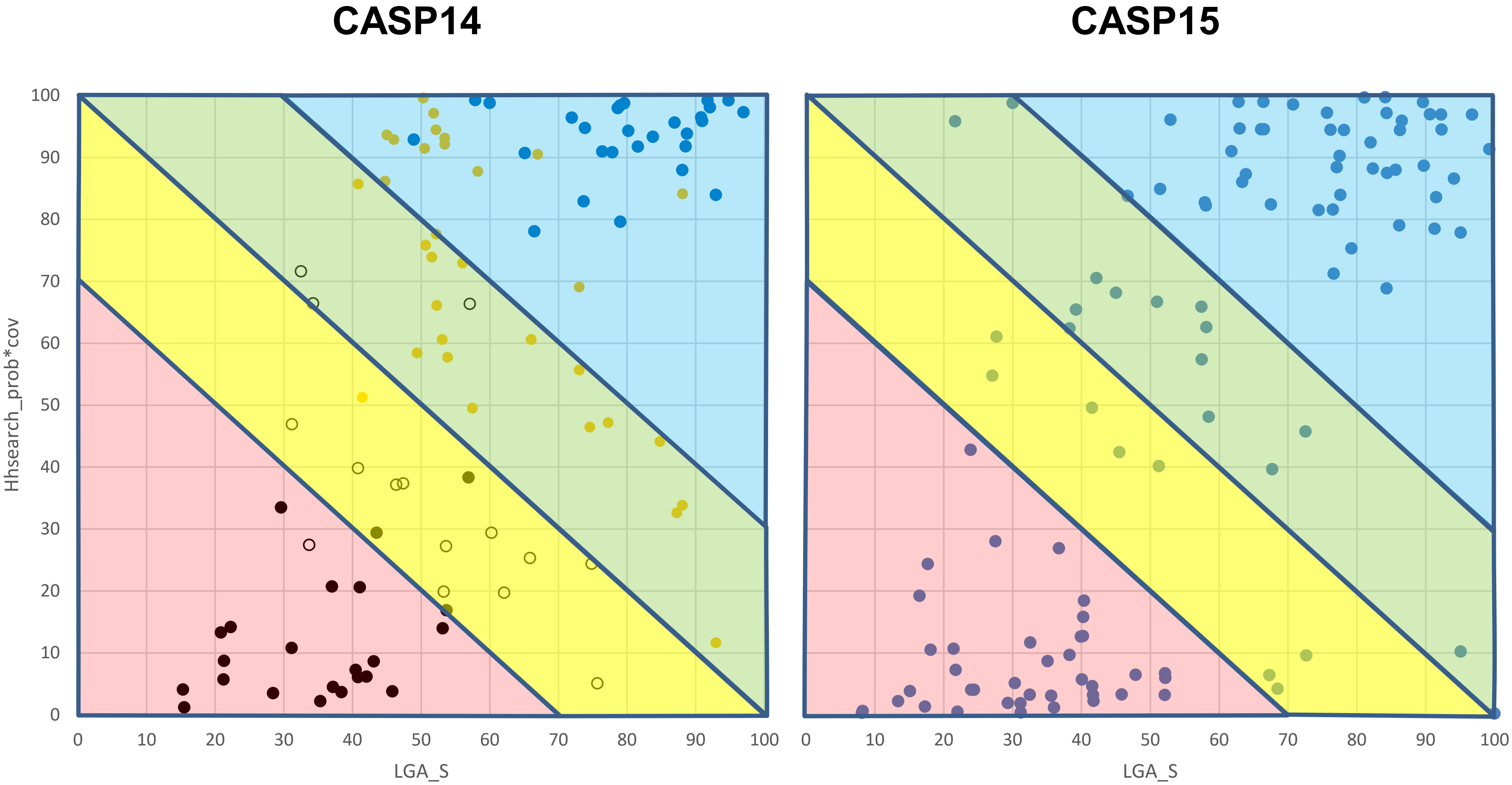
D2: 1302-2735

D3: 378-699,1223-1301

D4: 700-1222

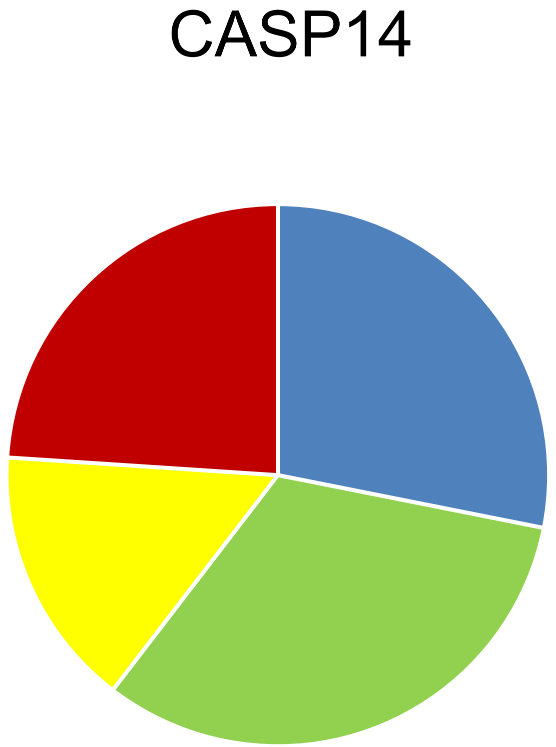


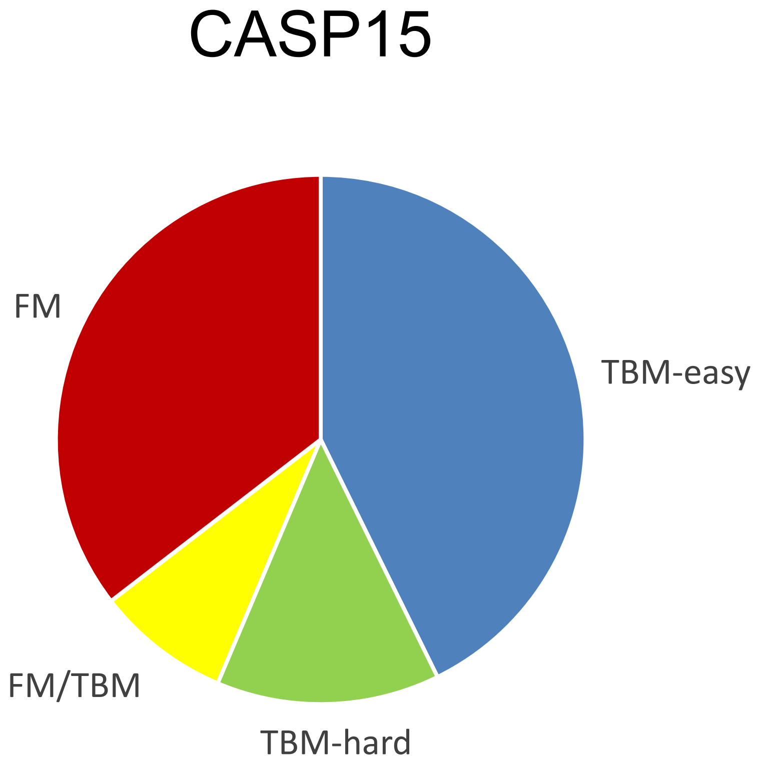
Domain classification

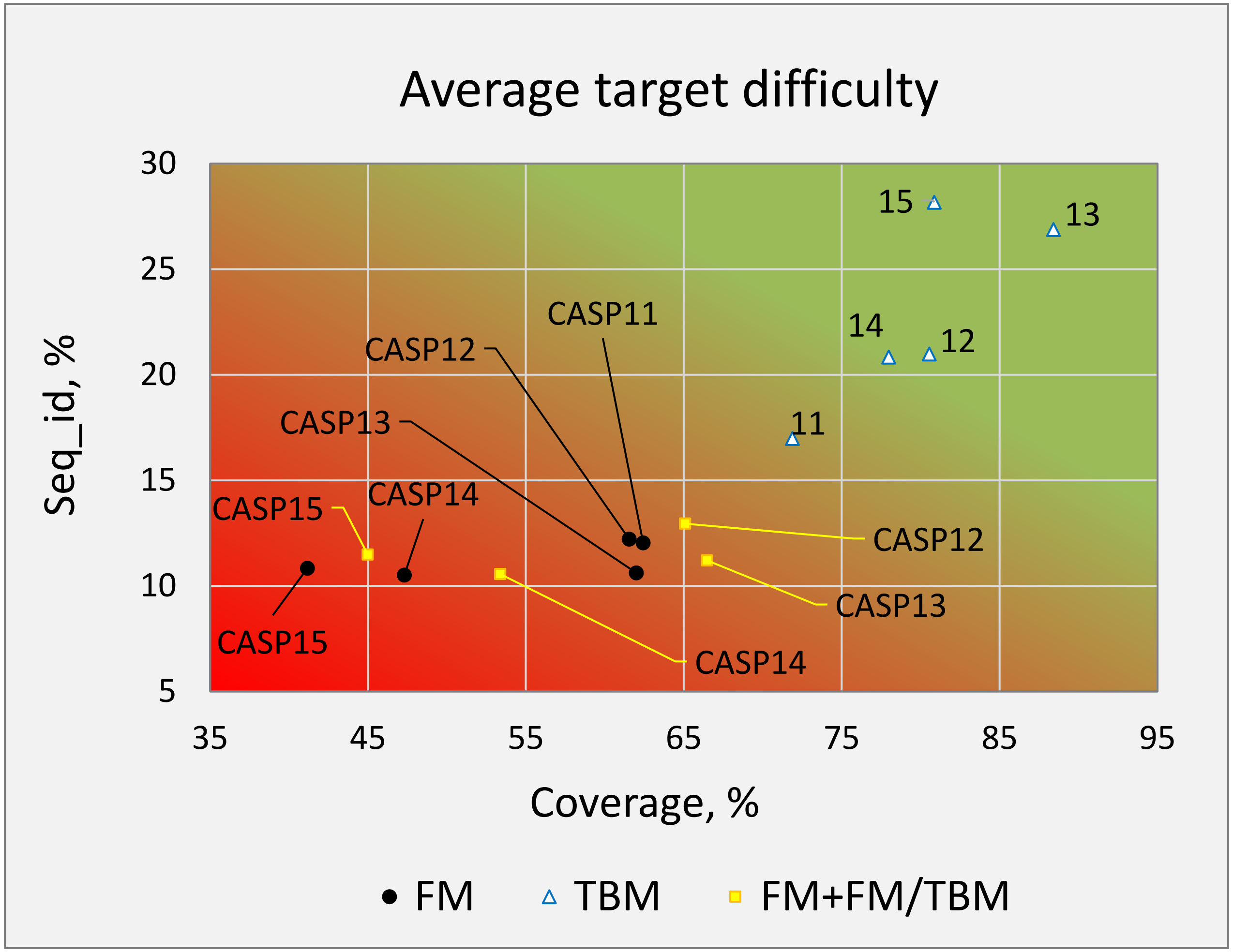


Domains (105 entries)

(difficulty categories)

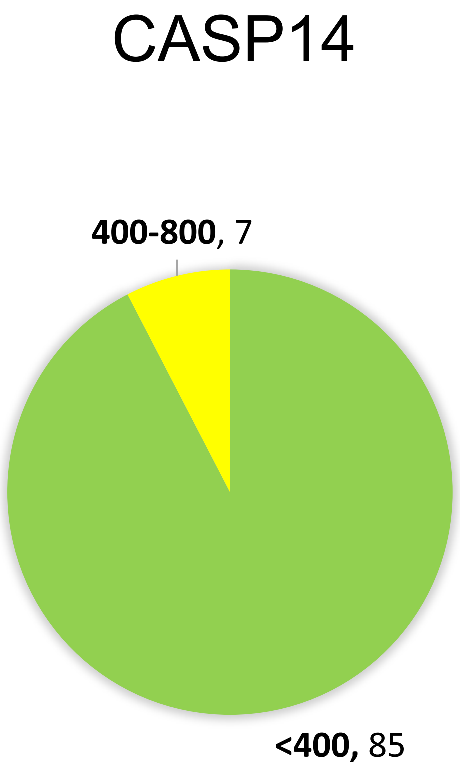


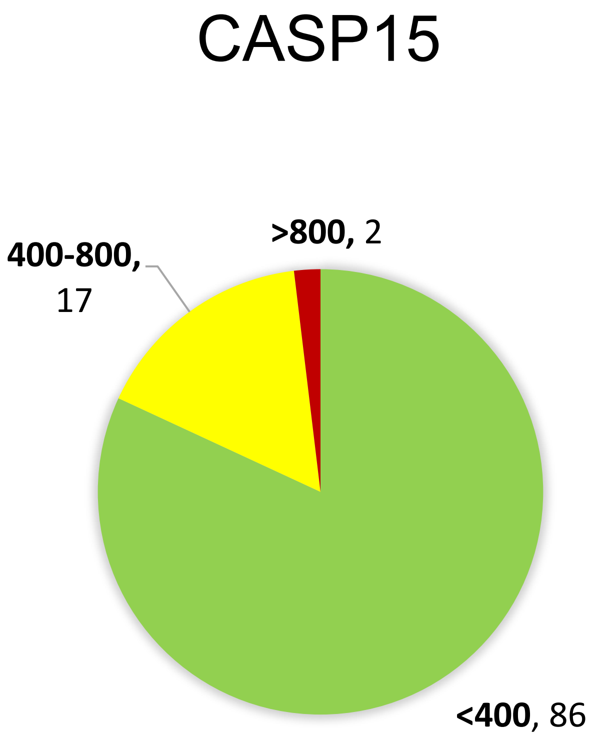



Average target difficulty
Seq_id, %
30
25
20
15
10
5
35
45
55
65
75
85
95
Coverage, %
15
13
14
12
11
CASP11
CASP12
CASP13
CASP14
CASP15
CASP15
CASP12
CASP13
CASP14
● FM
△ TBM
▫ FM+FM/TBM

Domains (105 entries)

(size)



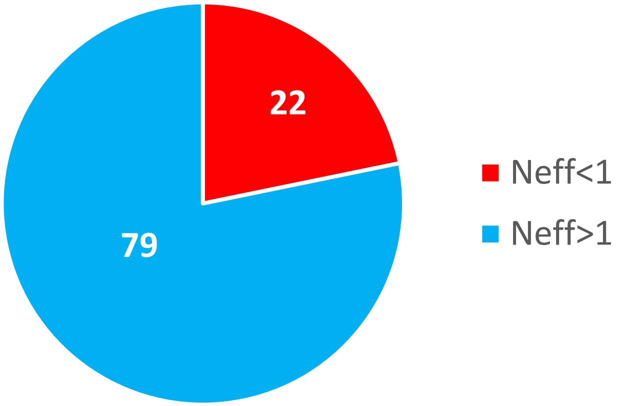


Availability of sequence relatives

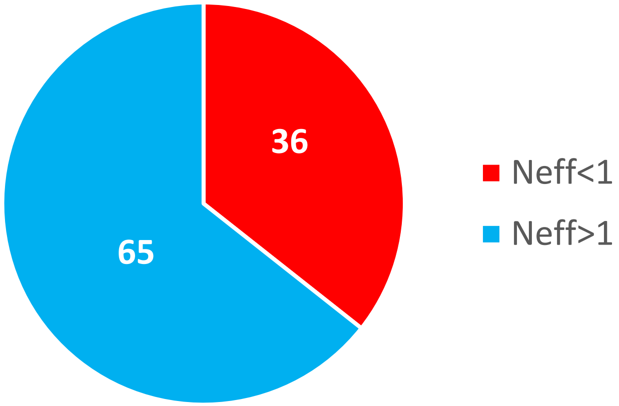
(Neff, CASP15)

	Andriy (Uniref)	Claudio (BF, Magnify)
T1122	0.00	0.00
T1130	0.01	0.01
T1131	0.01	0.01
T1125-D4	0.01	0.01
T1125-D1	0.01	0.01
T1125-D5	0.01	0.02
T1119	0.05	0.03
T1113	0.03	0.04
T1123-D1	0.02	0.05
T1125-D2	0.01	0.05
T1129s2	0.03	0.07
T1154-D1	0.02	0.09
T1178	0.10	0.11
T1173-D2	0.02	0.12
T1154	0.01	0.12
T1154-D2	0.03	0.14
T1125-D6	0.03	0.15
T1184	0.06	0.17
T1169-D4	0.20	0.21
T1125-D3	0.02	0.35
T1179	0.42	0.45
T1159	0.43	0.51
T1155	1.09	0.94

Neff (BFD, Magnify)
Claudio Mirabello

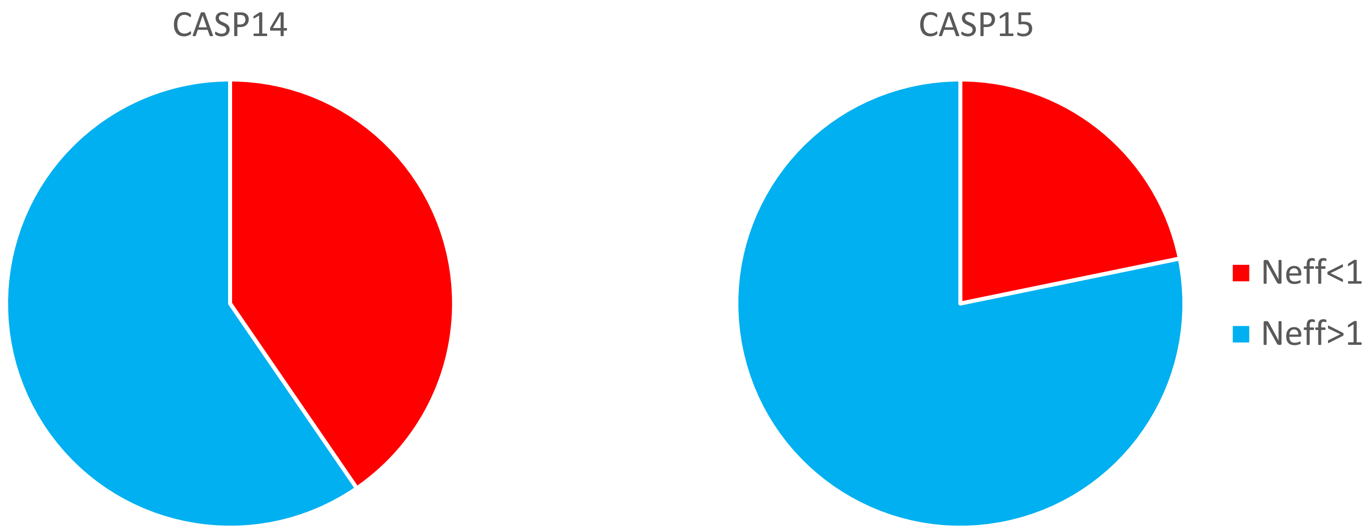


Neff (Uniref)
Andriy



	Andriy	Claudio
T1122	0.00	0.00
T1130	0.01	0.01
T1131	0.01	0.01
T1154	0.01	0.12
T1125-D5	0.01	0.02
T1125-D2	0.01	0.05
T1125-D4	0.01	0.01
T1125-D1	0.01	0.01
T1125-D3	0.02	0.35
T1154-D1	0.02	0.09
T1173-D2	0.02	0.12
T1123-D1	0.02	0.05
T1181-D2	0.03	8.28
T1154-D2	0.03	0.14
T1125-D6	0.03	0.15
T1113	0.03	0.04
T1129s2	0.03	0.07
T1119	0.05	0.03
T1184	0.06	0.17
T1181	0.09	8.02
T1178	0.10	0.11
T1145	0.19	5.14
T1169-D4	0.20	0.21
T1169-D1	0.25	3.47
T1173	0.26	13.57
T1145-D2	0.27	4.81
T1180	0.35	27.81
T1157s1	0.37	6.56
T1158	0.38	14.24
T1193	0.39	13.63
T1174-D2	0.40	3.69
T1175	0.41	9.53
T1179	0.42	0.45
T1159	0.43	0.51
T1120-D1	0.44	2.29
T1176	0.57	1.92
T1106s1	0.64	2.02
T1162	0.74	4.50
T1182	0.79	7.27
T1194	0.82	6.81
T1158-D2	0.84	24.34
T1165-D2	0.97	3.17

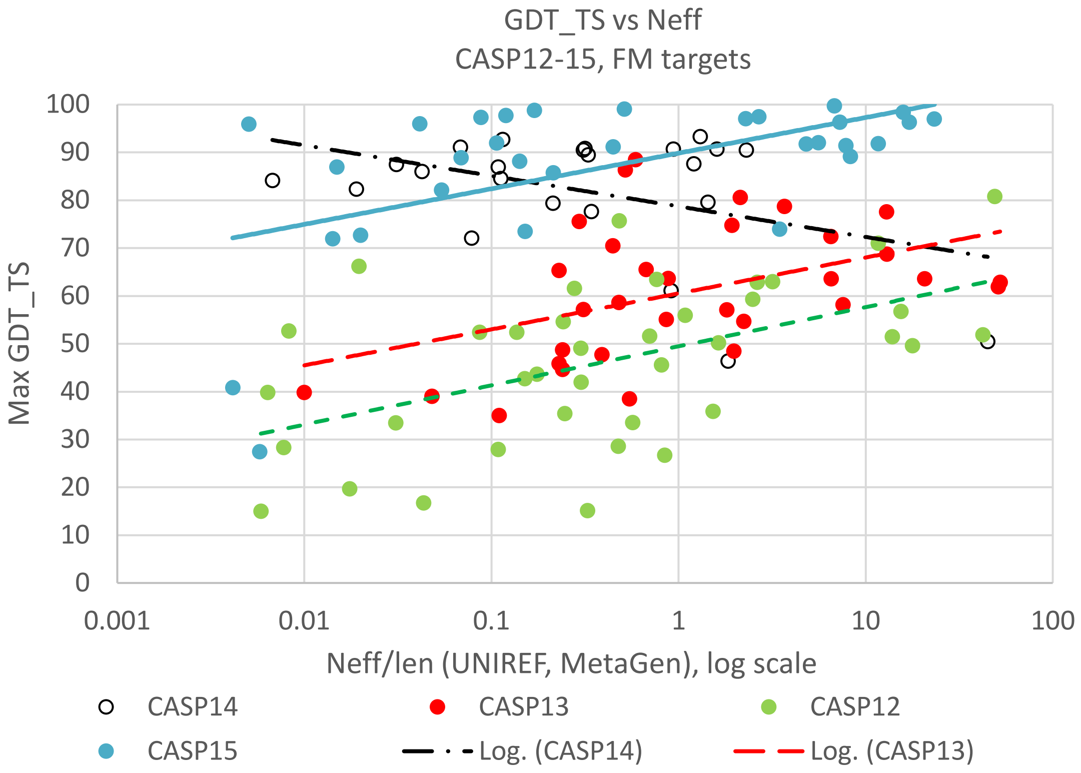
Availability of sequence relatives (Neff)



CASP15: 3 singletons; 11 domains < 10 seq

CASP14: 1 singleton; 14 domains < 10 seq

Performance vs MSA depth



THANKS