Protein structure prediction server: $(PS)^2$

Chih-Chieh Chen (陳志杰)

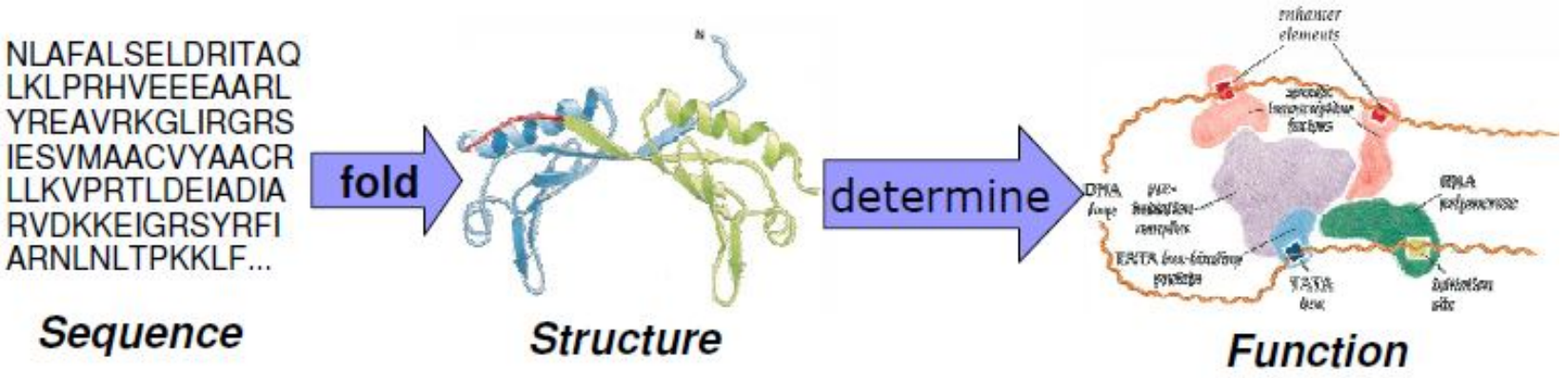
September 1, 2010

Outline

- Introduction
- Methods
  - Generation of the S2A2-matrix
  - Alignment scoring function
- Results and discussion
  - Alignment
  - Database search
  - Fold recognition
- Application
  - $(PS)^2$-v2: protein structure prediction server

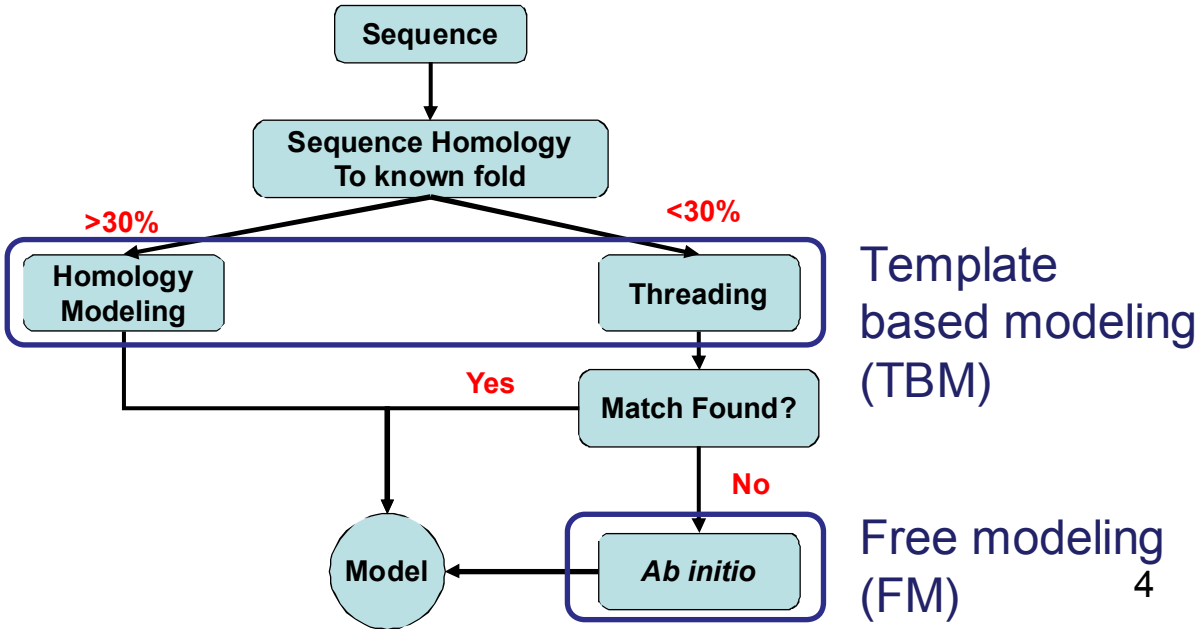
Why study protein structure?

- Proteins play crucial functional roles in all biological processes: enzymatic catalysis, signaling messengers …
- Function depends on 3D structure.
- Easy to obtain protein sequences, difficult to determine structure.

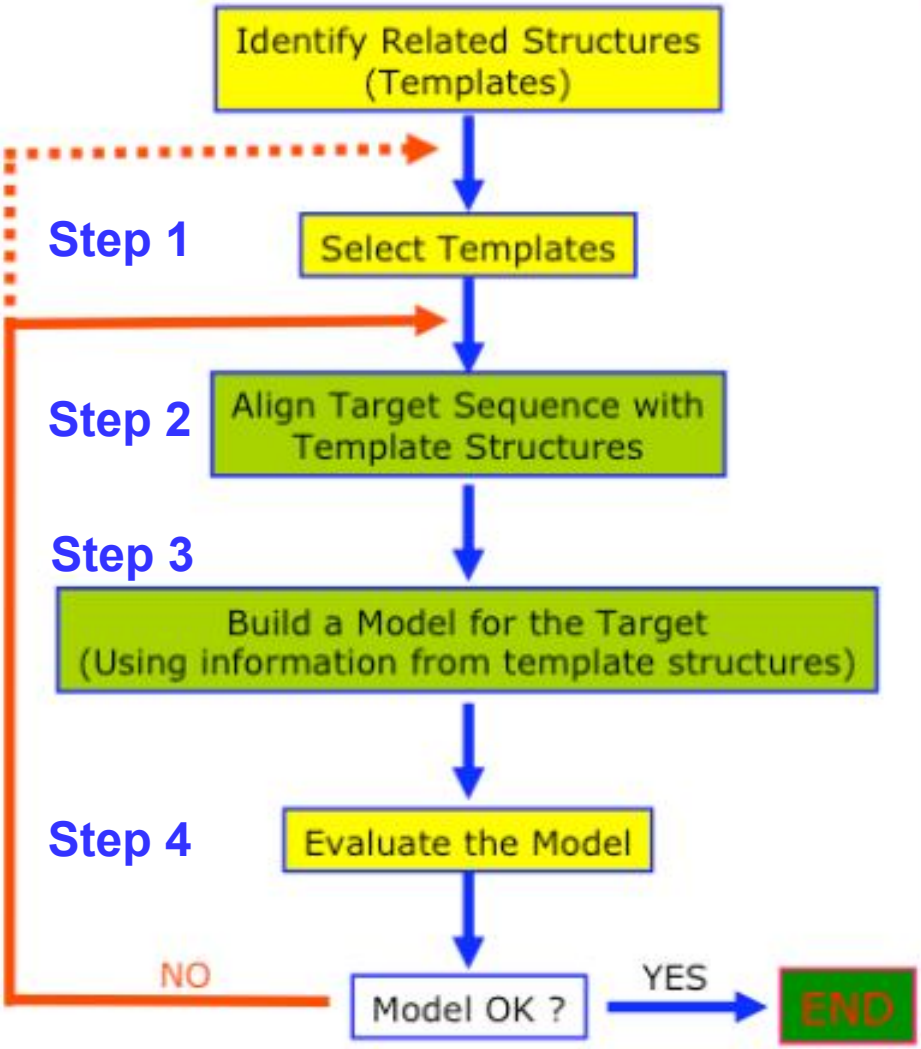


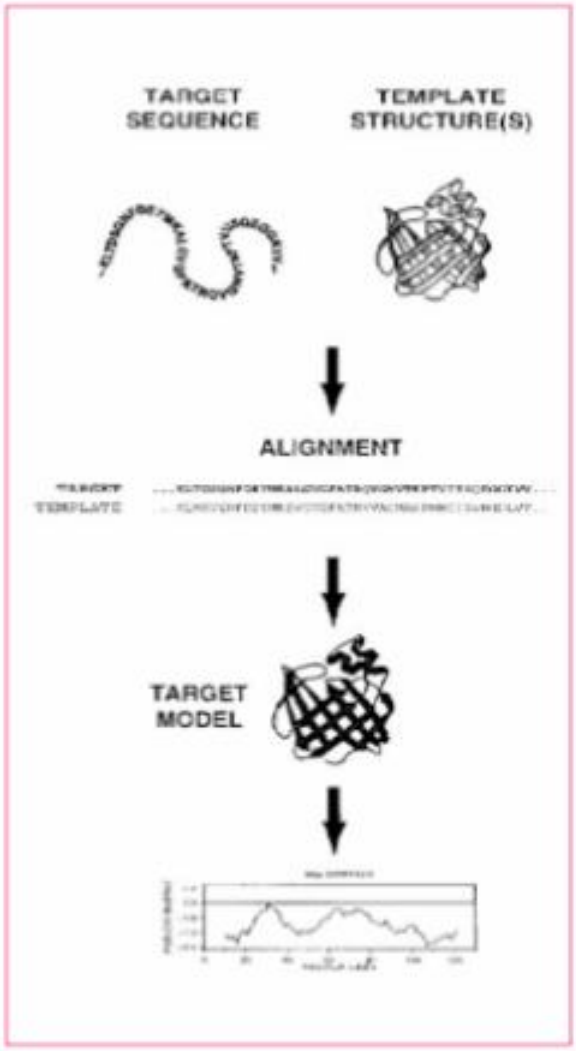
Introduction

- The primary sequence already contain all the information necessary to define 3D structure.
- The 3D protein structure can be predicted according to three main categories of methods (Rost & O'Donoghue, 1997): (1) **homology modeling**; (2) **fold recognition (threading)**; (3) ***ab initio* techniques**.



Structure prediction by TBM





Sequence based methods

- Sequence-to-sequence
  - BLAST (Altschul *et al.*, 1990)
  - FASTA, SSEARCH (Pearson, 1991)

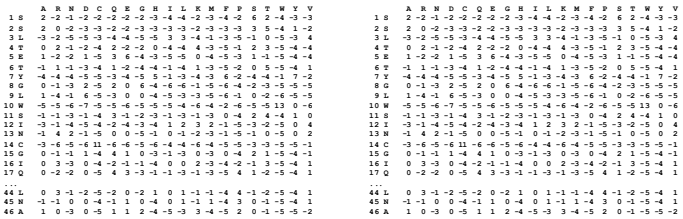
…NSKVRCLALMKGLE…
…QTWVKCIVLMRGLD…

- Sequence-to-profile
  - PSI-BLAST (Altschul *et al.*, 1997)
  - IMPALA (Schaffer *et al.*,1999)
  - HMMER (Durbin *et al.*, 1998)
  - SAM (Karplus *et al.*, 1998)

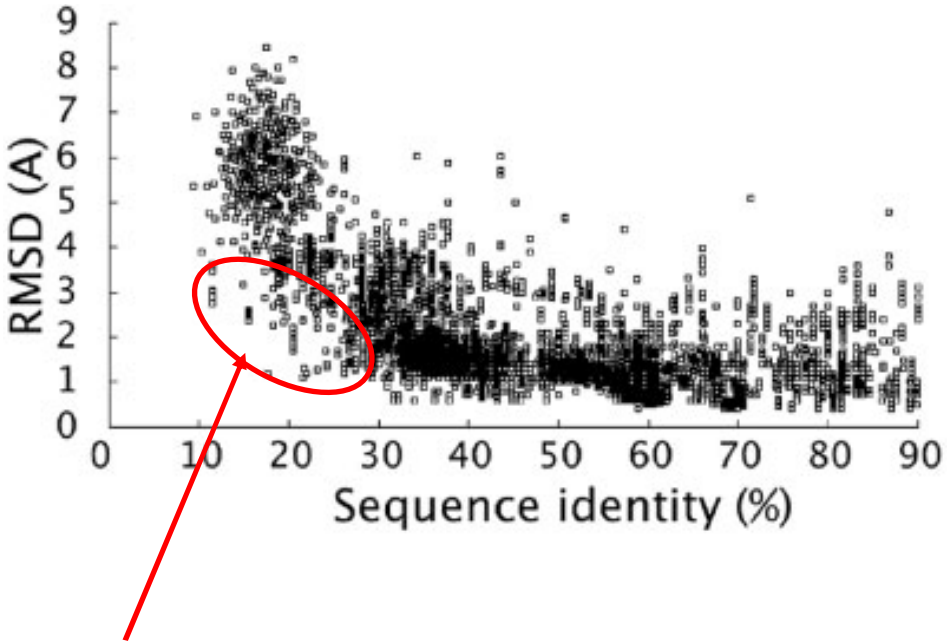
…NSKVRCLALMKGLE…

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	S	2	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	-2	-3	-4	-2	6	2	-4	-3	-3
2	S	2	0	-2	-3	-3	-2	-2	-3	-3	-3	-3	-2	-3	-3	-3	3	5	-4	1	-2
3	L	-3	-2	-5	-5	-3	-4	-4	-5	-5	3	3	-4	-1	-3	-5	-1	0	-5	-3	4
4	T	0	2	-1	-2	-4	2	-2	-2	0	-4	-4	4	-3	-5	-1	2	3	-5	-4	-4
5	E	1	-2	-2	1	-5	3	6	-4	-3	-5	-5	0	-4	-5	-3	1	-1	-5	-4	-4
6	T	-1	1	-1	-3	-4	1	-2	-4	-4	-1	-4	1	-3	-5	-2	0	5	-5	-4	1
7	Y	-4	-4	-4	-5	-5	-3	-4	-5	5	-1	-3	-4	-3	6	-2	-4	-4	-1	7	-2
8	G	0	-1	-3	2	-5	2	0	6	-4	-6	-6	-1	-5	-6	-4	-2	-3	-5	-5	-5
9	L	1	-4	-1	6	-5	-3	0	0	-4	-5	-3	-3	-5	-6	-1	0	-2	-6	-5	-5
10	W	-5	-5	-6	-7	-5	-5	-6	-5	-5	-5	-4	-6	-4	-2	-6	-5	-5	13	0	-6
11	S	-1	-1	-3	-1	-4	3	-1	-2	-3	-1	-3	-1	-3	0	-4	2	4	-4	1	0
12	I	-3	-1	-4	-5	-4	-2	-4	-3	-4	1	2	3	-2	-1	-5	-3	-2	-5	0	4
13	N	-1	4	2	-1	-5	0	0	-5	-1	0	-1	-2	-3	-1	-5	-1	0	-5	0	2
14	C	-3	-6	-5	-6	11	-6	-6	-5	-6	-4	-4	-6	-4	-5	-5	-3	-3	-5	-5	-1
15	G	0	-1	-1	1	-4	4	1	0	-3	-1	-3	0	-3	0	-4	2	1	-5	-4	-1
16	I	0	3	-3	0	-4	-2	-1	-1	-4	0	0	2	-3	-4	-2	-1	3	-5	-4	1
17	Q	0	-2	-2	0	-5	4	3	-3	-1	-1	-3	-1	-3	-5	4	1	-2	-5	-4	1
...																					
44	L	0	3	-1	-2	-5	-2	0	-2	1	0	1	-1	-1	-4	4	-1	-2	-5	-4	1
45	N	-1	-1	0	0	-4	-1	1	0	-4	0	1	-1	1	-4	3	0	-1	-5	-4	1
46	A	1	0	-3	0	-5	1	1	2	-4	-5	-3	3	-4	-5	2	0	-1	-5	-5	-2

- Profile-to-profile
  - *prof_sim* (Yona & Levitt, 2002)
  - COMPASS (Sadreyev & Grishin, 2003)



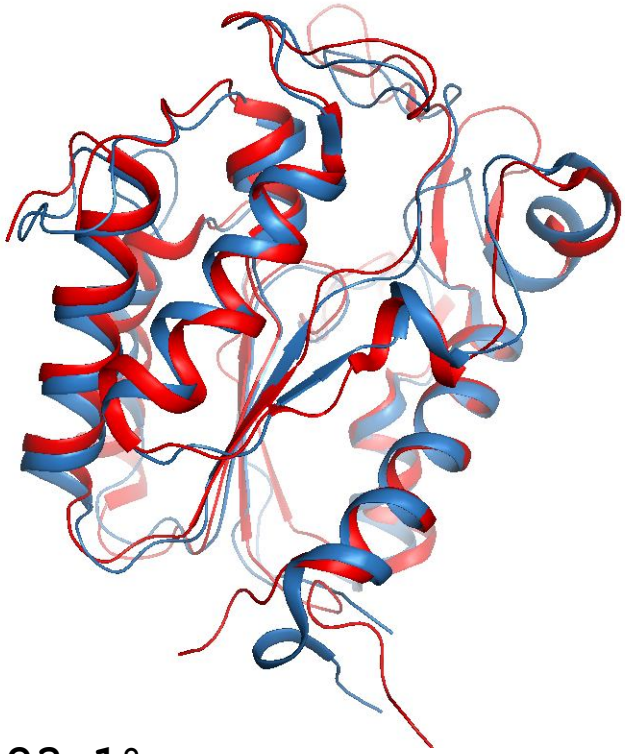
Relationship between sequence and structure



remote homologs

(Chuang *et al*., 2003)

Including secondary structure information



3dc7A <-> 3bzwA,
RMSD = 1.7Å,
Z-score = 6.6

3dc7A <-> 3bzwA, Sequence identity = 23.1%

```
CCCCCCCCEEEEEECCCCCCC--cccCCCHHHHHHHHHHCCcccEEEECCCCcCCCHHHH
SNGHVSFKRPAWLGDSITANN--GLATVHYHDILAADWDVERSDNLGISGSTIGSRYDAM
       K+  ++GDSIT  N  G     Y D L     +        + G + G ++D +
IQHPWQGKKVGYIGDSITDPNCYGDNIKKYWDFLKEWLGI----TPFVYGIS-GRQWDDV
CCCCCCCCEEEEEECCCCCCCCcchhhCCCHHHHHHHHHCC----EEEECCCC-CCCHHHH
```

Including secondary structure information

- Clustering: reducing 20 amino acids to several classes
  - 3D-1D substitution matrix (Rice & Eisenberg, 1997)

Table 2. The seven residue classes of the H3P2 matrix and their solvent accessibility boundaries

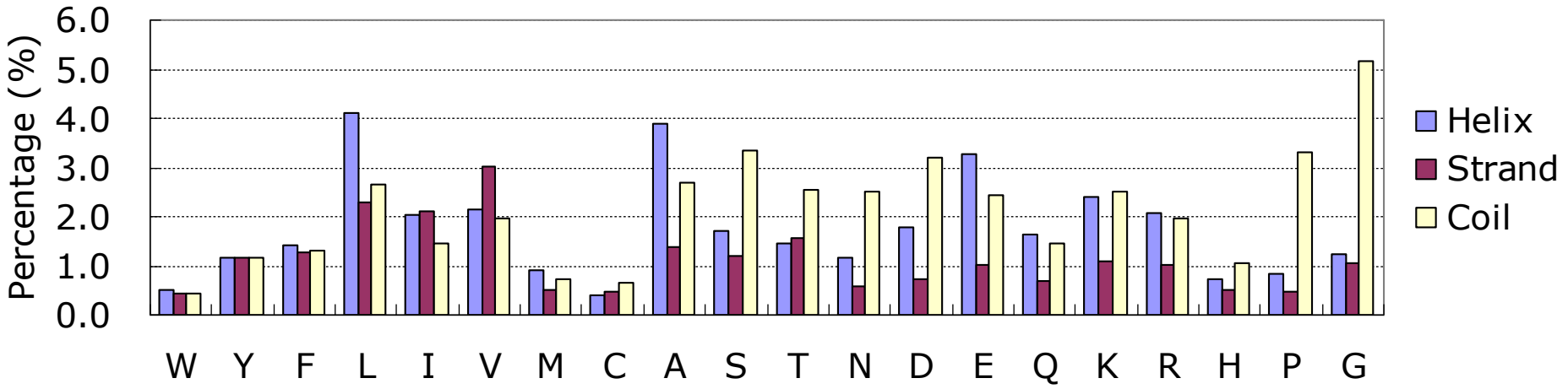
Symbol	Property	Residues					Boundary
r_c	Cysteine	Cys					0.99
r_w	Tryptophan	Trp					0.95
r_b	Basic	Arg	Lys				0.67
r_a	Aromatic	Tyr	Phe				0.95
r_h	Hydrophobic	Ile	Leu	Met	Val		0.98
r_s	Small	Ala	Gly	Ser	Thr	Pro	0.84
r_s	Polar	Asp	Glu	Gln	Asn	His	0.74

The clustering of the 20 amino acids into the seven residue classes used in this study. Clustering was done according to substitution values in the PAM250 matrix. Column 1 gives the symbolic name used in the H3P2 matrix (Table 1), column 2 gives a brief description of a common property of residues in each class, and column 3 lists the amino acids belonging to each residue class. Column 4 gives the buried/exposed boundary for each residue class in units of fractional residue area buried. A value of 0.99 means that if only 1% of a residue in that class is exposed to solvent it is assigned to an exposed class.

Including secondary structure information

- Separation: linear contribution to the substitution score
  - 3D-PSSM (Kelley *et al.*, 2000) — $(+1/-1)$
  - AA/SS (Wallqvist *et al.*, 2000) — (3×3)
  - Matras (Kawabata & Nishigawa, 2000) — $S_{i,j} = \alpha M_{i,j}^{seq} + \beta M_{i,j}^{ss}$
  - *two-track* HMMs (Karchin *et al.*, 2003) — *Secondary structure profile comparison*
  - HAMP hybrid profiles (Tang *et al.*, 2003) — *Secondary structure profile comparison*
  - SPARKS (Zhou & Zhou, 2004)
  - SSALN (Qiu & Elber, 2006)
  - SP^3 (Zhou & Zhou, 2005)
  - SP^4 (Liu *et al.*, 2007)

The secondary structure and the amino acid distribution in a position are strongly dependent on each other.



20 AA × 3 SS = 60 **RS** letters

Substitution matrix derived from structural alignments

- 674 structure pairs from SCOP 1.65
  - RMSD < 3.5Å
  - Sequence identity < 40%
- Using DSSP program to assign secondary structure
  - (H, G, I) → H
  - (E, B) → E
  - (T, S, blank) → C
- Computing a log odds matrix

Calculation of the S2A2-matrix

- The scores of S2A2-matrix are define as:

$$S_{ij} = \log_2(q_{ij} / e_{ij})$$

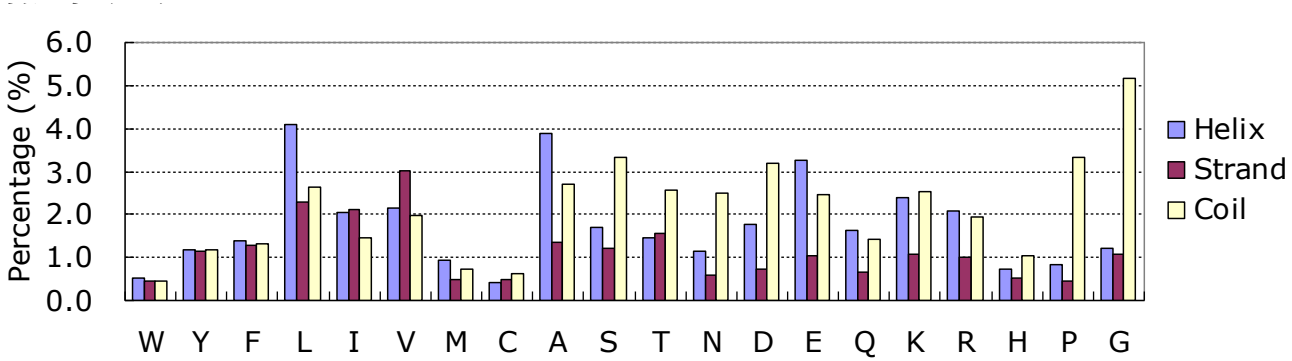
$$q_{ij} = f_{ij} / \sum_{i=1}^{60} \sum_{j=1}^{i} f_{ij}$$

$$e_{ij} = \begin{cases} p_i p_j & if \quad i = j \\ 2 p_i p_j & if \quad i \neq j \end{cases}, \qquad p_i = q_{ii} + \sum_{k \neq i}^{60} q_{ik} / 2$$

  - S_{ij} is the S2A2-matrix score of *i*, *j* pair
  - q_{ij} observed probability of *i*, *j* pair
  - e_{ij} expected probability of *i*, *j* pair

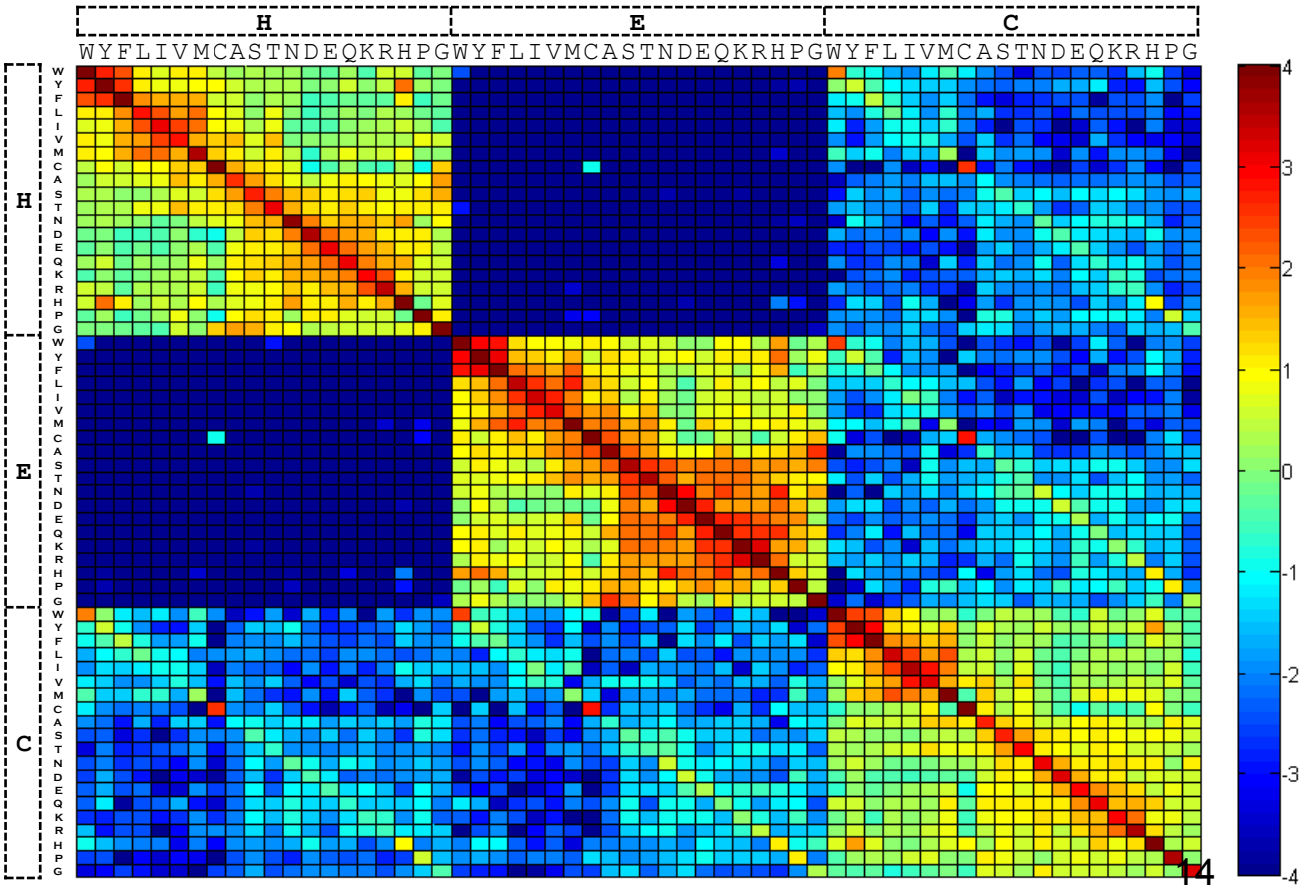
(Henikoff *et al*., 1992)

General properties of the S2A2-matrix



- **Properties**

1. Mean = -2
2. Max = 6.2
3. E/E > H/H > C/C

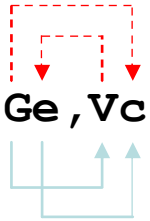


Correlation between S2A2 & BLOSUM62

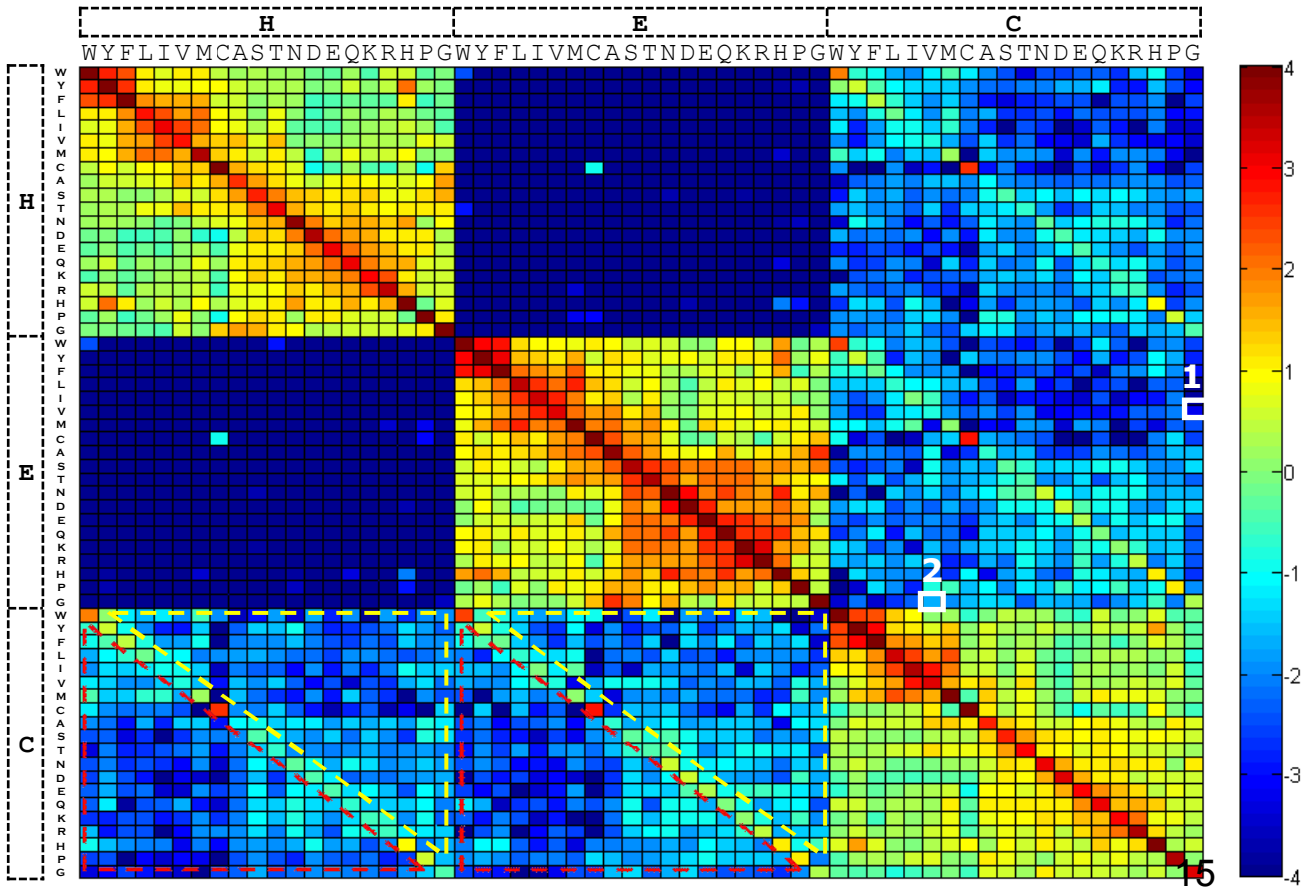
	BLOSUM	H->H	E->E	C->C	H->C	E->C	H->E
BLOSUM	1.000						
H->H	0.941	1.000					
E->E	0.941	0.935	1.000				
C->C	0.934	0.921	0.898	1.000			
H->C	0.691	0.704	0.659	0.698	1.000		
E->C	0.688	0.675	0.685	0.684	0.598	1.000	
H->E	0.240	0.206	0.214	0.255	0.242	0.265	1.000

- **Asymmetry**

1. *S*(Ve,Gc) = -3.2



2. *S*(Ge,Vc) = -1.7



corr = 0.358 *corr* = 0.287

Position Specific Scoring Matrix (PSSM)

20 amino acid types

Sequence length

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	S	2	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	-2	-3	-4	-2	6	2	-4	-3	-3
2	S	2	0	-2	-3	-3	-2	-2	-3	-3	-3	-3	-2	-3	-3	-3	3	5	-4	1	-2
3	L	-3	-2	-5	-5	-3	-4	-4	-5	-5	3	3	-4	-1	-3	-5	-1	0	-5	-3	4
4	T	0	2	-1	-2	-4	2	-2	-2	0	-4	-4	4	-3	-5	-1	2	3	-5	-4	-4
5	E	1	-2	-2	1	-5	3	6	-4	-3	-5	-5	0	-4	-5	-3	1	-1	-5	-4	-4
6	T	-1	1	-1	-3	-4	1	-2	-4	-4	-1	-4	1	-3	-5	-2	0	5	-5	-4	1
7	Y	-4	-4	-4	-5	-5	-3	-4	-5	5	-1	-3	-4	-3	6	-2	-4	-4	-1	7	-2
8	G	0	-1	-3	2	-5	2	0	6	-4	-6	-6	-1	-5	-6	-4	-2	-3	-5	-5	-5
9	L	1	-4	-1	6	-5	-3	0	0	-4	-5	-3	-3	-5	-6	-1	0	-2	-6	-5	-5
10	W	-5	-5	-6	-7	-5	-5	-6	-5	-5	-5	-4	-6	-4	-2	-6	-5	-5	13	0	-6
11	S	-1	-1	-3	-1	-4	3	-1	-2	-3	-1	-3	-1	-3	0	-4	2	4	-4	1	0
12	I	-3	-1	-4	-5	-4	-2	-4	-3	-4	1	2	3	2	-1	-5	-3	-2	-5	0	4
13	N	-1	4	2	-1	-5	0	0	-5	-1	0	-1	-2	-3	-1	-5	-1	0	-5	0	2
14	C	-3	-6	-5	-6	11	-6	-6	-5	-6	-4	-4	-6	-4	-5	-5	-3	-3	-5	-5	-1
15	G	0	-1	-1	1	-4	4	1	0	-3	-1	-3	0	-3	0	-4	2	1	-5	-4	-1
16	I	0	3	-3	0	-4	-2	-1	-1	-4	0	0	2	-3	-4	-2	-1	3	-5	-4	1
17	Q	0	-2	-2	0	-5	4	3	-3	-1	-1	-3	-1	-3	-5	4	1	-2	-5	-4	1
...																					
44	L	0	3	-1	-2	-5	-2	0	-2	1	0	1	-1	-1	-4	4	-1	-2	-5	-4	1
45	N	-1	-1	0	0	-4	-1	1	0	-4	0	1	-1	1	-4	3	0	-1	-5	-4	1
46	A	1	0	-3	0	-5	1	1	2	-4	-5	-3	3	-4	-5	2	0	-1	-5	-5	-2

```
QUERY    1   SSLTETYGLWSINCGIQE-G---KK-V--C-FMHRQEVNDQ-N-RVVVAMSVVL---N-A 46
2011621 31   STLQETYQDWTVSCQSQK-E---TS-I--C-VMRQDQSSTQ-T-GQRVLTAELR---N-V 76
2012088 30   STLTETYQDWSISCQAQK-E---ST-N--C-IMNQMQNSSQ-TGQRVLTVELRN---V-A 76
2534888 42   SSLQETYQDWSLACQ-SA-P---QK-V--C-VISQQQVQPN-G-QRVLAIELRS---G-G 86
209527  31   STLQETYQDWTVSCQSQK-D---TT-A--C-VMRQEQSSAQ-A-GQRVLTAELR---NVA 77
2926303 43   ASVSETYGDWTVDCRLVD-R---RK-Q--C-LLSQVQGNKE-TGRRVYAIELTP---P-A 89
2257946 70   SAVSEVYGDWTINCTRES-A---AR-R--C-ALSQAQGEAQ-TGRRRFSIELRP---P-E 116
1370088 76   -TLTERFKAWTVTC-VEA-E---TR-T--C-RMTQELV--Q-Q-KTGQRLSALL---V-E 111
2534887 5    TTLSETYEAWTVQ-SVKA-GEGARR-M--C-RMSQELIQPE-T-RQRVLLFAIT---K-G 52
```

Scoring and alignment method of ProS2A2

- Scoring function: combination of **S2A2-matrix** and **PSSM**

$$S_{i,j} = \max \begin{cases} S_{i-1,j-1} + w^{stru}(i,j) w^{S2A2} S2A2(i,j) + (1 - w^{S2A2}) PSSM_{query}(i,j) \\ S_{i-1,j} + w^{gap}(i,j) w^{S2A2} g^{S2A2} + (1 - w^{S2A2}) g^{pssm} \\ S_{i,j-1} + w^{gap}(i,j) w^{S2A2} g^{S2A2} + (1 - w^{S2A2}) g^{pssm} \\ 0 \end{cases}$$

  - *i, j* are the indexes of the query and template sequences
  - *S2A2*(*i*, *j*) is the S2A2-matrix score
  - *PSSM*(*i*, *j*) is the PSSM score
  - w^{S2A2} is the weight of the S2A2-matrix
  - $w^{stru}(i, j)$ is the structure dependent weight for S2A2-matrix
  - $w^{gap}(i, j)$ is the structure dependent gap penalty for S2A2-matrix

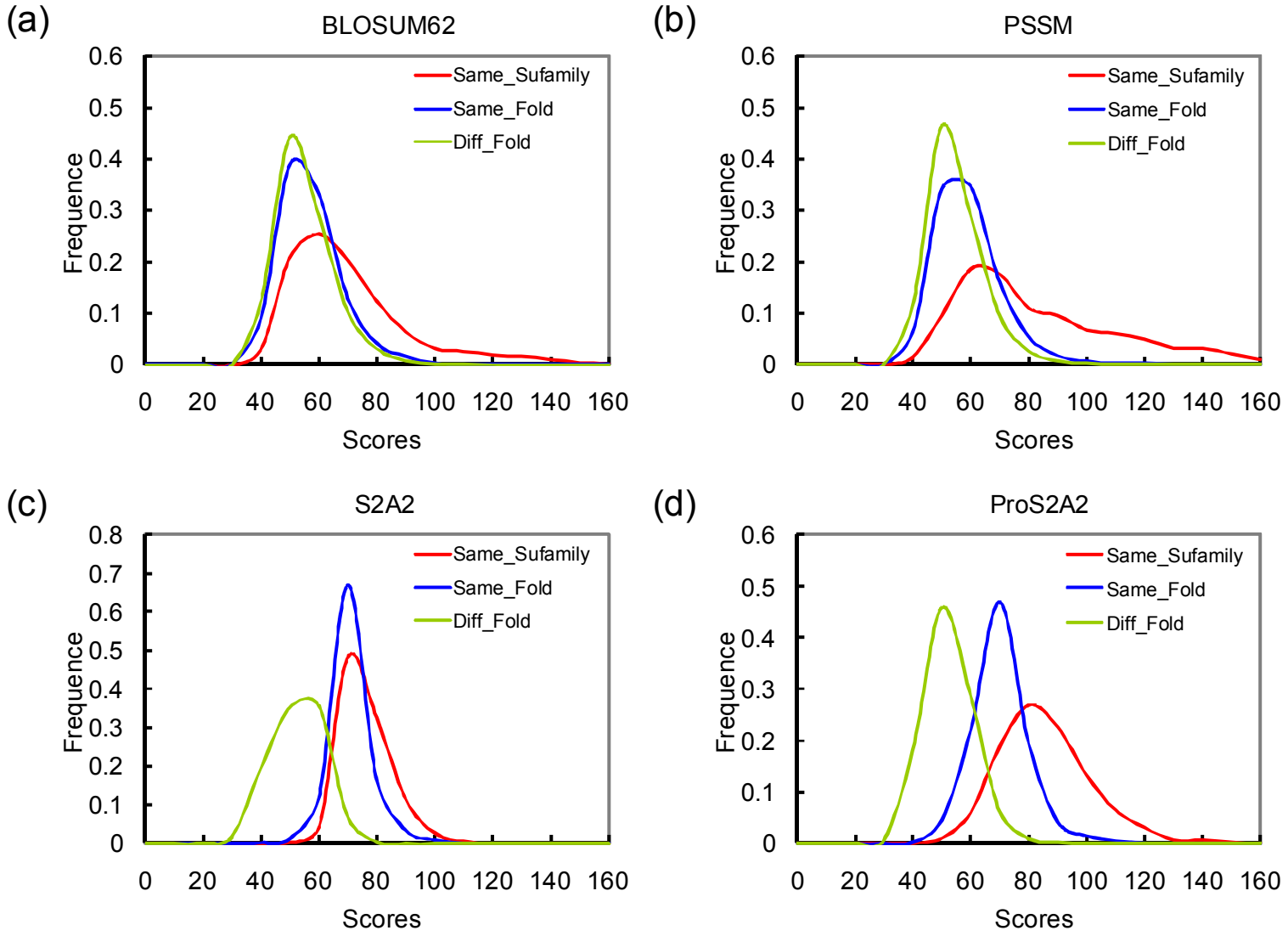
Optimization of parameters using SALIGN set

- 200 structure pairs
  - RMSD < 3.5 Å
  - Sequence identity < 20%
- Reference alignments
  - CE program
- Performance evaluation
  - CE overlap

(Marti-Renom *et al.*, 2004)

Alignment accuracy results on the SALIGN set

Method	CE overlap (%)
BLOSUM62 (20×20)	30.4
PSSM	45.8
BLOSUM62+SS	47.1
S2A2 (60×60)	49.3
ProS2A2 (S2A2+PSSM)	**56.3**



Distribution of similarity scores for different scoring systems.

ProSup benchmark

- 127 structure pairs
  - Sequence identity < 30%
- Reference alignments
  - ProSup program
- Performance evaluation
  - $T_{\mathbf{c}\text{orrect}}$, $T_{\mathbf{m}\text{issed}}$, $T_{\mathbf{i}\text{ncorrect}}$, σ_0

(Domingues *et al*., 2000)

Comparing ProS2A2 with other methods for sequence alignment accuracy on the ProSup benchmark

Method	T_c [e]	T_m [e]	T_i [e]	σ_0 [e]
BL62+SS [a]	7983	1315	6527	48.6
S2A2 [a]	8732	947	7198	53.4
ProS2A2 [a]	**9470**	**868**	**6998**	**58.7**
SSALN [b]	9256	1115	7245	58.3
SPARKS [c]	-	-	-	57.2
prof_sim	8009	4505	3142	43.6
PSI-BLAST	6733	4938	3452	36.4
FASTA [d]	5340	3003	7452	31.4

[a] This work.

[b] Results from Qiu and Elber.

[c] Results from Zhou and Zhou.

[d] Results from Domingues et al..

[e] T_c and T_m are total numbers of correctly aligned and missed residue pairs, respectively; T_i is the total number of incorrect aligned pairs; σ_0 is the average percentage of correctly aligned residues.

(a)

PSSM (Only sequence)

```
1lmb3 <-> 1pou       Shift = 0.60          s0 = 58.49
6 QLEDARRLKAIYEKKKNELGLSQ----ESVADKMG--MGQSGVGALFNGINALNAYN********* 56
  HHHHHHHHHHHHHHHHHCCCCh----hhhhhhhc--cCHHHHHHHHcccccchhh********* 
  |||||||||||||||||||||||      ----         |||||||||    ++++
  000000000000000000000000      4444         000000000    4444
  CHHHHHHHHHHHHHHHHHCCCcchhhhhhhhhhcccccCCHHHHHHH----ccccc********* 
1 DLEELEQFAKTFKQRRIKLGFTQGDVGLAMGKLYGNDFSQTTISRFE----ALNLSF********* 53
```

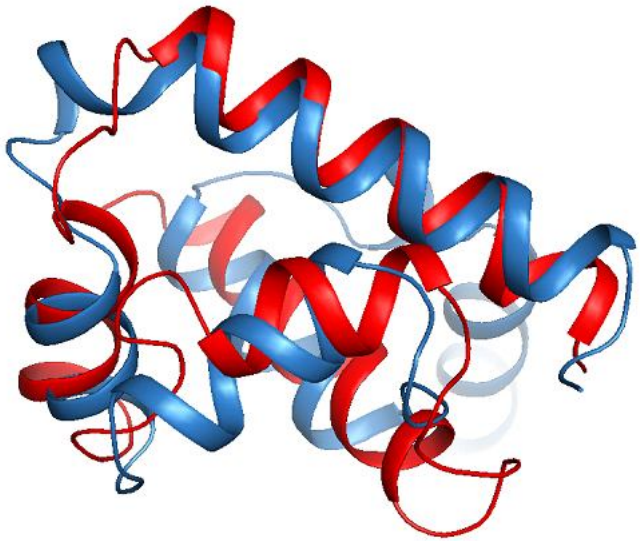
S2A2-matrix

```
1lmb3 <-> 1pou       Shift = 0.67          s0 = 60.78
7 LEDARRLKAIYEKKKNELGLS----QESVADKMG--MGQSGVGALFN-GINALNAYNAALLAKILKVSVEEFS 72
  HHHHHHHHHHHHHHHHHCCCC----hhhhhhhhc--cCHHHHHHHHC-ccccchhhhhhhhhhhcchhhcc
  |||||||||||||||||||||      ----         |||||||||| --        ++++++++    ++
  000000000000000000000      4444         0000000000 11        11111111    44
  HHHHHHHHHHHHHHHHHHCCCcchhhhhhhhhhcccccCCHHHHHHHCccccccchhhhhhhhhhh---hhhcc
2 LEELEQFAKTFKQRRIKLGFTQGDVGLAMGKLYGNDFSQTTISRFEALNLSFKNMCKLKPLLEKW---LNDAE 71
```

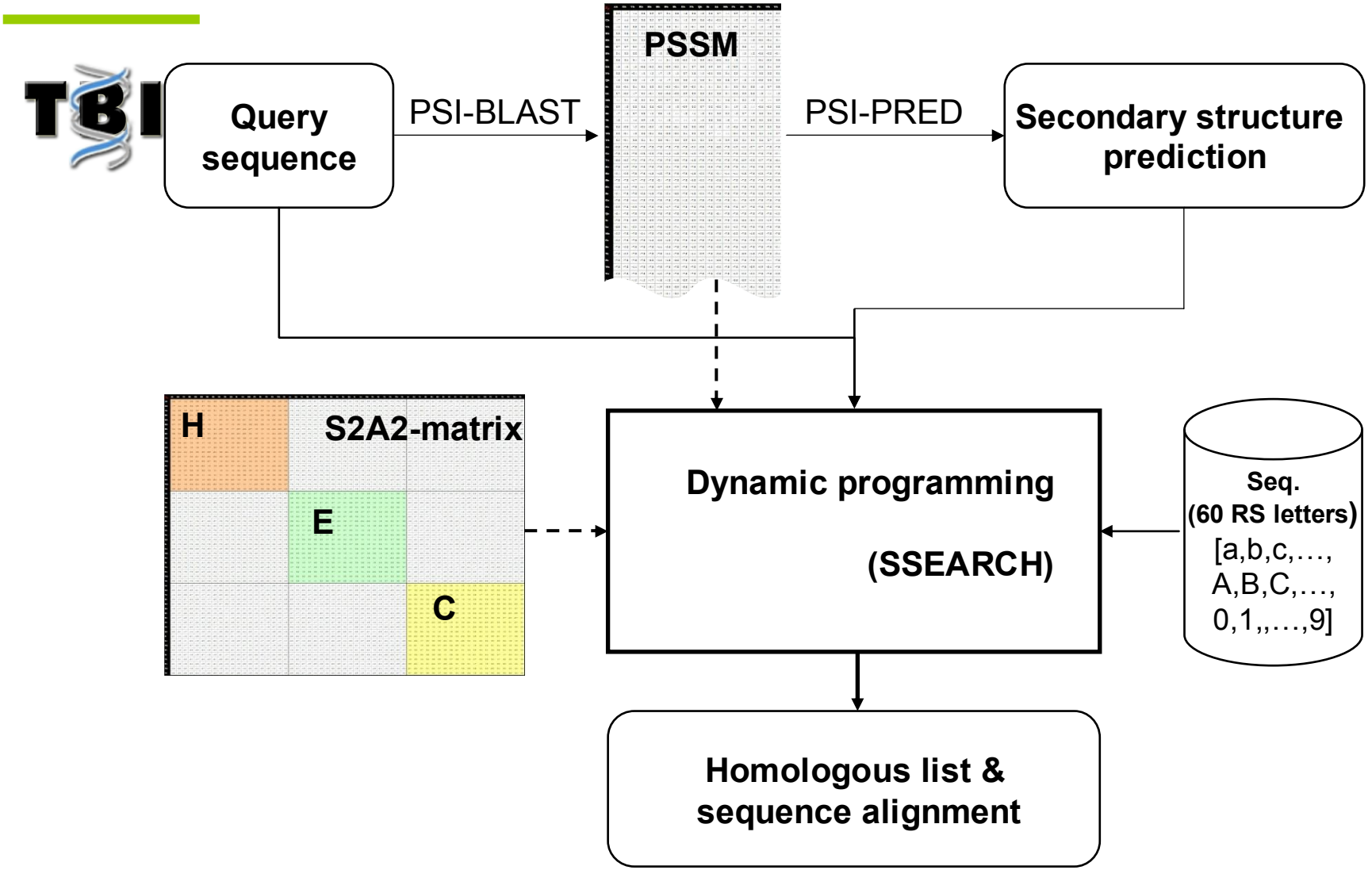
S2A2-matrix + PSSM (ProS2A2)

```
1lmb3 <-> 1pou       Shift = 0.39          s0 = 74.51
7 LEDARRLKAIYEKKKNELGLS----QESVADKMG--MGQSGVGALFN-GINALNAYNAA-LLAKILK 65
  HHHHHHHHHHHHHHHHHCCCC----hhhhhhhhc--cCHHHHHHHHC-ccccchhhhhh+HHHHHHC
  |||||||||||||||||||||      ----         |||||||||| --        ++ ||||||| 
  000000000000000000000      4444         0000000000 11        11 00000000
  HHHHHHHHHHHHHHHHHHCCCcchhhhhhhhhhcccccCCHHHHHHHCccccccchhhhhhhHHHHHHHH
2 LEELEQFAKTFKQRRIKLGFTQGDVGLAMGKLYGNDFSQTTISRFEALNLSFKNMCKLKPLLEKWLN 68
```

(b)



Example of the effect of S2A2-matrix and PSSM on our alignment method.



A diagram of the flow chart in the PorS2A2 system.

Calculation of *E*-value

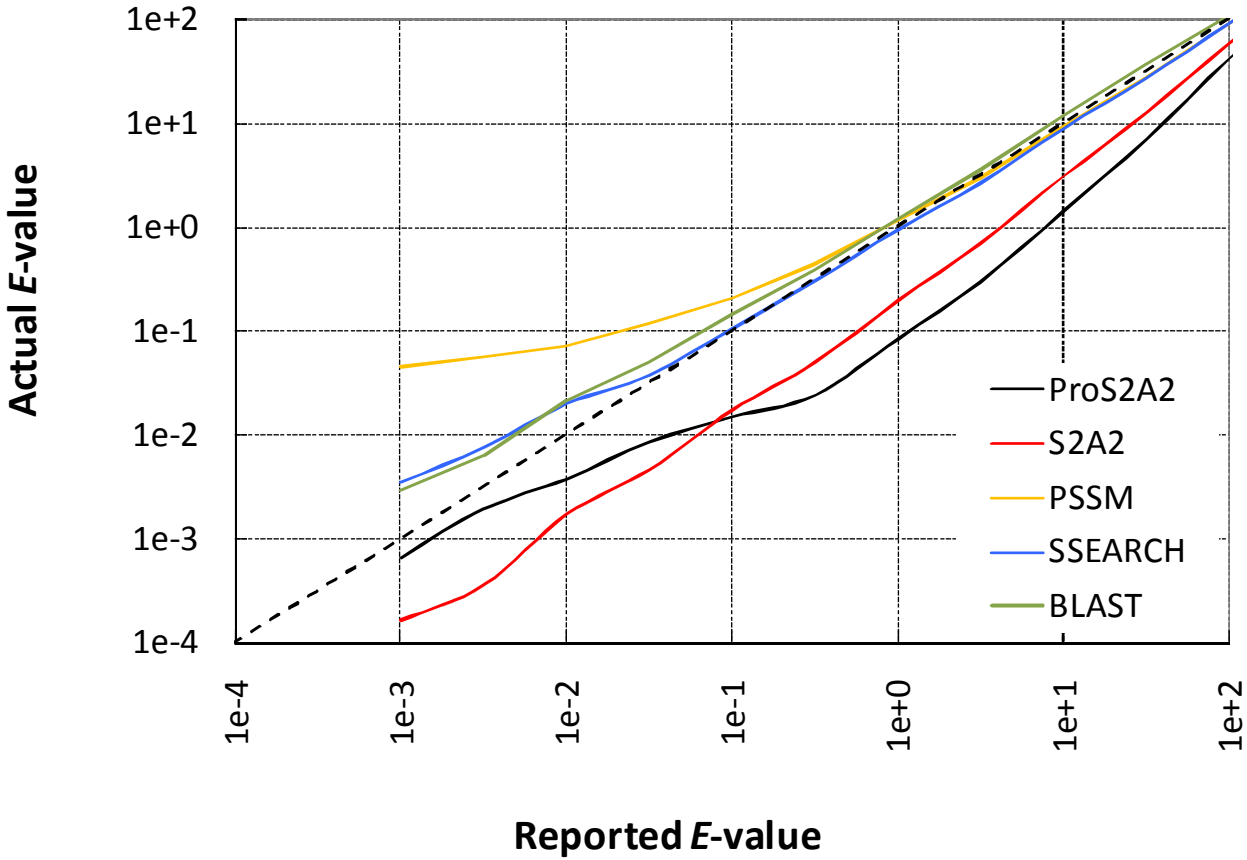
- Rescaling the row scores by calculation of *E*-value

$$E = Kmne^{-\lambda S}$$

  - *m*, *n* are lengths of the two sequences
  - *S* is the score of the optimal alignment
  - λ , *K* are statistical parameters

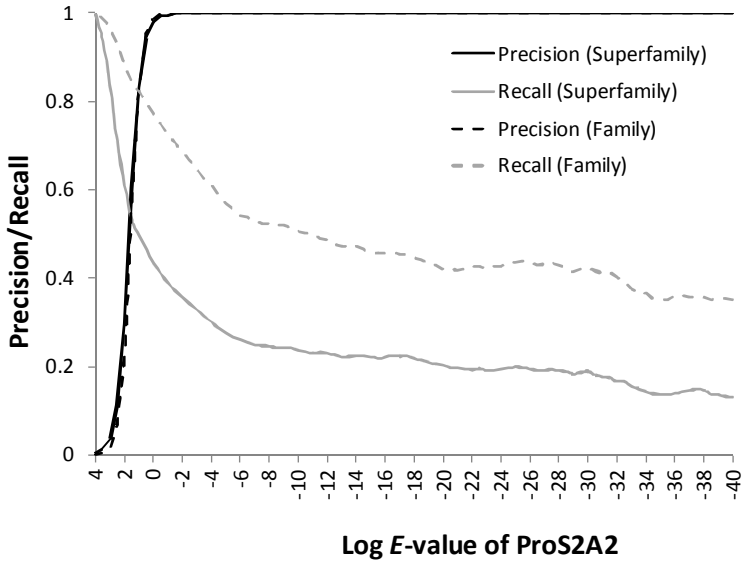
SCOP-1926 benchmark

- 1926 proteins
  - Sequence identity < 40% (in SCOP 1.73, not in SCOP 1.71)
- Search database
  - SCOP 1.71 (11709 sequences)
- Evaluation
  - Database search
  - Superfamily assignment

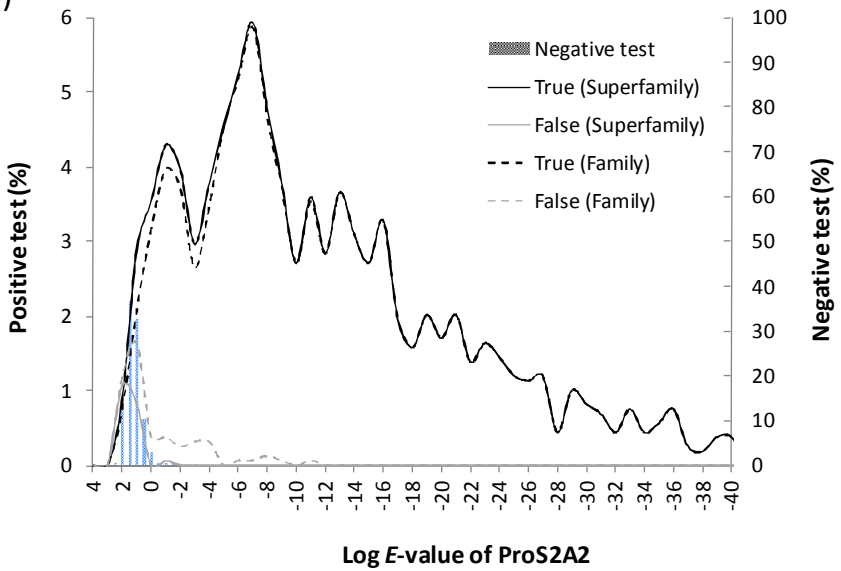


Actual versus reported *E*-value on the SCOP-1926 dataset.

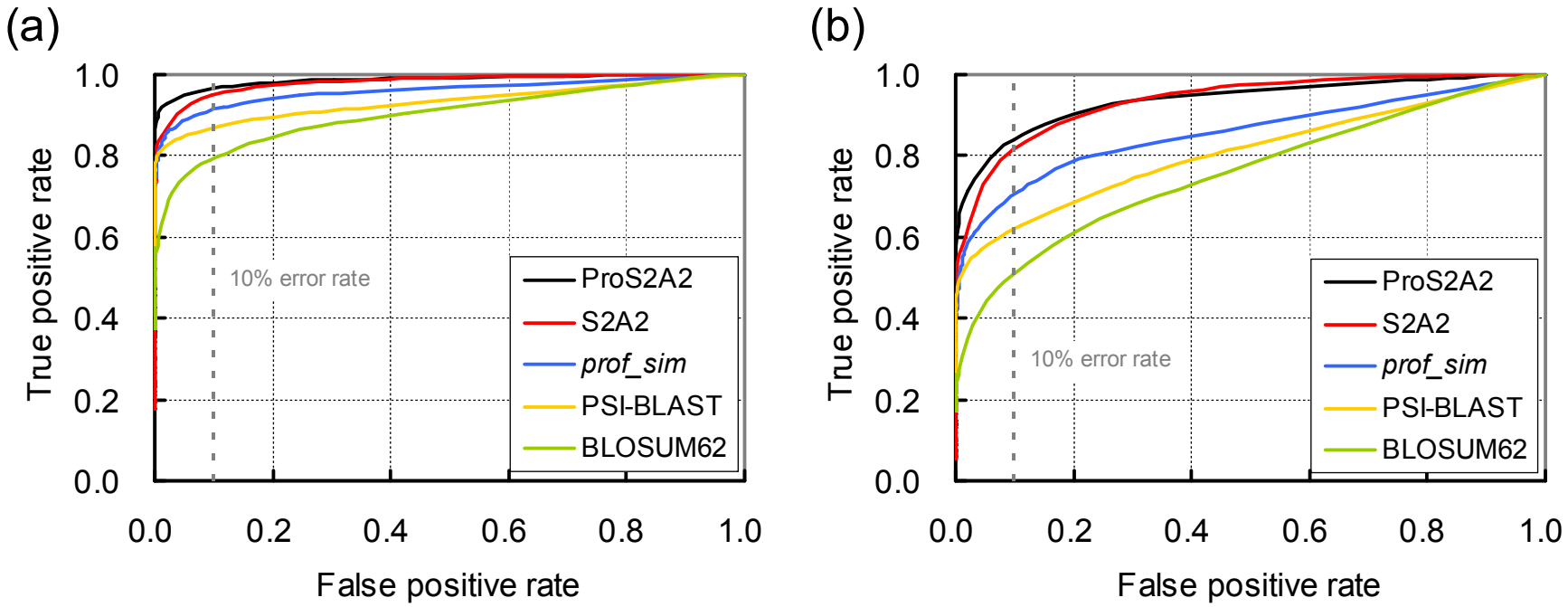
(a)



(b)



ProS2A2 search performance with *E*-values on the query set SCOP-1926.



Comparison of the ProS2A2 with other methods based on SCOP-1926 benchmark. (a) Family, (b) Superfamily.

ProS2A2 results using beta-hydroxyacyl-ACP dehydratase FabZ as the query

SCOP code	Protein name	SCOP family name	ProS2A2 *E*-value	PSI-BLAST *E*-value	RMSD (Å)	Sequence identity (%)
d1u1za_	(3R)-hydroxymyristoyl ACP dehydrase FabZ	FabZ-like	4.2e-15	8.0e-50	2.07	47.3
d1mkaa_	beta-Hydroxydecanol thiol ester dehydrase	beta-Hydroxydecanol thiol ester dehydrase	4.8e-9	7.0e-29	1.94	21.5
d1q6wa_	Monoamine oxidase regulatory protein	MaoC-like	2.6e-3	2.0e-3	4.63	22.5
d1ixla_	Hypothetical protein PH1136	PaaI/YdiI-like	6.4e-3	6.0e-3	3.84	23.0
d1vpma_	Acyl-CoA hydrolase BH0798	4HBT-like	8.3e-3	1.5	3.05	16.3
d1j1ya_	Phenylacetic acid degradation protein PaaI	PaaI/YdiI-like	0.016	-	3.04	19.6
d1sc0a_	Hypothetical protein HI1161	PaaI/YdiI-like	0.019	-	3.33	17.0
d1psua_	Phenylacetic acid degradation protein PaaI	PaaI/YdiI-like	0.023	1.2	4.66	20.6
d1iq6a_	(R)-specific enoyl-CoA hydratase	MaoC-like	0.025	1.0e-3	3.53	20.9
d1nnga_	Putative acyl-coa thioester hydrolase HI0827	4HBT-like	0.039	-	4.24	17.8
d1q4ta_	4-hydroxybenzoyl CoA thioesterase	PaaI/YdiI-like	0.058	-	3.35	17.9
d1njka_	Hypothetical protein YbaW	4HBT-like	0.11	0.12	3.80	15.5
d1vh5a_	Hypothetical protein YdiI	PaaI/YdiI-like	0.22	-	3.51	18.8
d1vh9a_	Hypothetical protein YbdB	PaaI/YdiI-like	0.36	-	3.56	16.7
d1t82a_	Putative thioesterase SO4397	PaaI/YdiI-like	0.43	-	3.47	15.9
d1lo7a_	4-hydroxybenzoyl-CoA thioesterase	4HBT-like	0.56	-	3.63	22.5
d1s5ua_	Hypothetical protein YbgC	4HBT-like	0.67	-	2.98	15.2
d1pn2a2	2-enoyl-coa hydratase domain of multifunctional peroxisomal hydratase-dehydrogenase-epimerase	MaoC-like	0.9	0.69	4.22	21.4
d1sh8a_	Hypothetical protein PA5026	PaaI/YdiI-like	1.7	-	4.13	20.2
d1c8ua2	Thioesterase II (TesB)	Acyl-CoA thioesterase	2.3	-	5.34	20.8

Lindahl benchmark

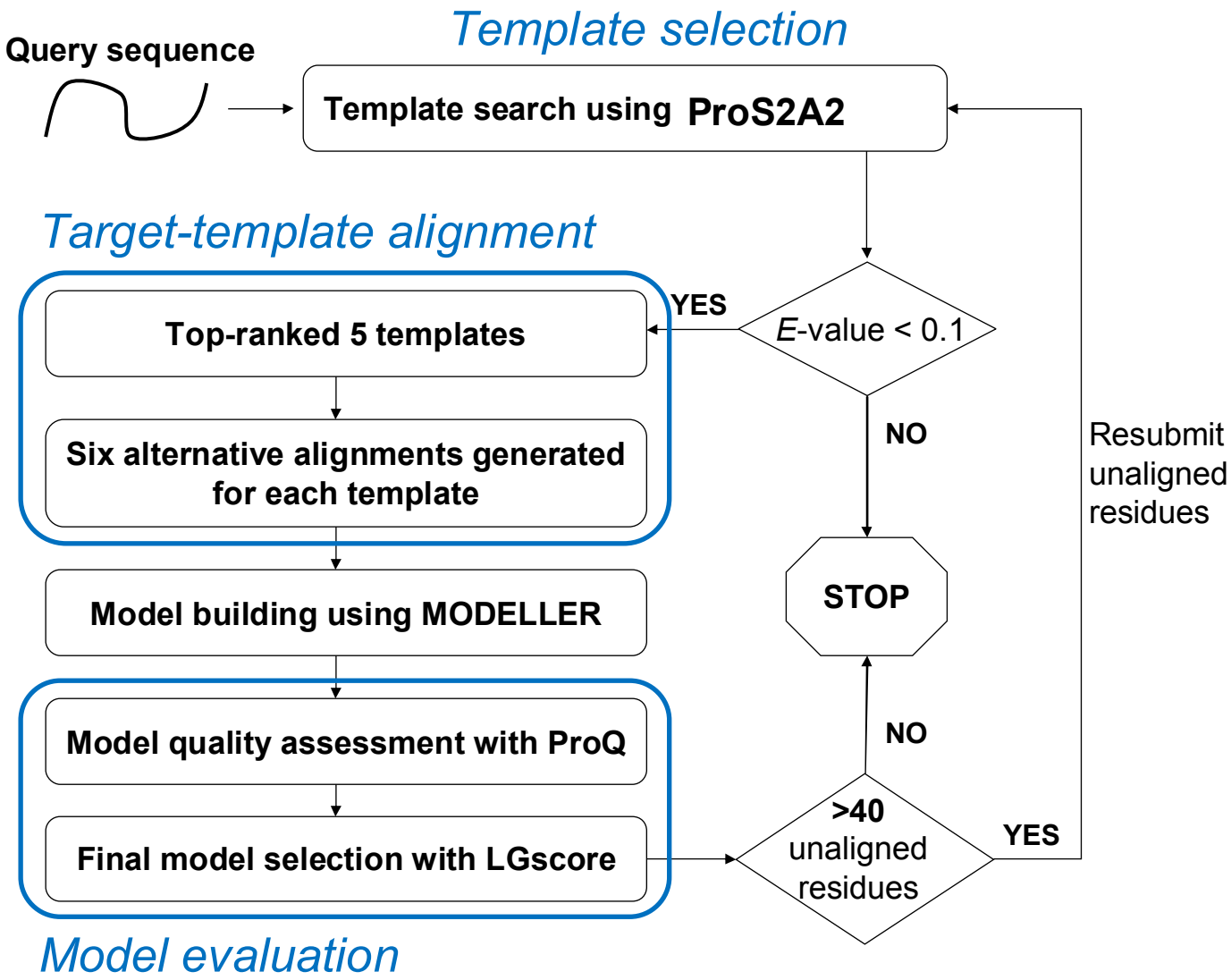
- 976 proteins
  - Sequence identity < 40%
- Possible correct hits
  - Family level (555)
  - Superfamily level (434)
  - Fold level (321)
- Evaluation
  - Fold recognition

(Lindahl & Elofsson, 2000)

Performance of fold recognition on the Lindahl benchmark

Methods	Family (%)		Superfamily (%)		Fold (%)	
	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5
S2A2	77.1	85.1	43.8	63.1	26.5	50.8
ProS2A2	82.2	88.8	56.7	**75.6**	27.1	**54.5**
BLAST	64.7	69.9	18.2	29.7	5.3	14.0
PSI-BLAST	74.4	79.5	38.5	49.1	4.4	14.6
HMMER	67.6	73.5	20.7	31.3	4.4	14.6
SAMT98	70.1	75.4	28.3	38.9	3.4	18.7
prof_sim	80.7	86.5	50.9	61.3	22.1	39.6
FUGUE	82.2	85.8	41.9	53.2	12.5	26.8
RAPTOR	84.8	87.1	47.0	60.0	31.3	54.2
PROSPECT II	84.1	88.2	52.6	64.8	27.7	50.3
SPARKS	81.6	88.1	52.5	69.1	24.3	47.7
FOLDpro	85.0	89.9	55.5	70.0	26.5	48.3
SP [3]	81.6	86.8	55.3	67.7	28.7	47.4
SP [4]	80.9	86.3	57.8	68.9	30.8	53.6

prof_sim (Yona & Levitt, 2002), FUGUE (Shi *et al.*, 2001), RAPTOR (Xu, 2003), PROSPECT II (Kim *et al.*, 2003), SPARKS (Zhou & Zhou, 2004), FOLDpro (Cheng & Baldi, 2006), SP^4 (Liu *et al.*, 2007)



The framework of the $(PS)^2$-v2 server for protein structure prediction.

The essential differences of $(PS)^2$-original, $(PS)^2$-CASP8 and $(PS)^2$-v2

Steps	$(PS)^2$-original	$(PS)^2$-CASP8	$(PS)^2$-v2
1. Template search	Consensus of PSI-BLAST and IMPALA	ProS2A2 with a self-developed aligned tool using dynamic programming	ProS2A2 with a modified SSEARCH program
2. Target-template alignment	Consensus of PSI-BLAST, IMPALA and T-coffee	ProS2A2 with a self-developed aligned tool using dynamic programming	ProS2A2 with a modified SSEARCH program
3. Template	Single template	Single template	Multiple templates
4. Model building	MODELLER with single model	MODELLER with single model	MODELLER with multiple models
5. Model evaluation	PROCHECK	PROCHECK	ProQ

(Chen *et al.*, 2006; Chen *et al.*, 2009)

$(PS)^2$ - consensus alignment

Input: target and template sequences

Output: target-template aligned sequences

Step 1: Initial all entries of the aligned matrix to 0. Align target and template sequences using PSI-BLAST, IMPALA, and T-Coffee.

Step 2: Sum aligned scores of these three alignments for each position with different scoring weights.

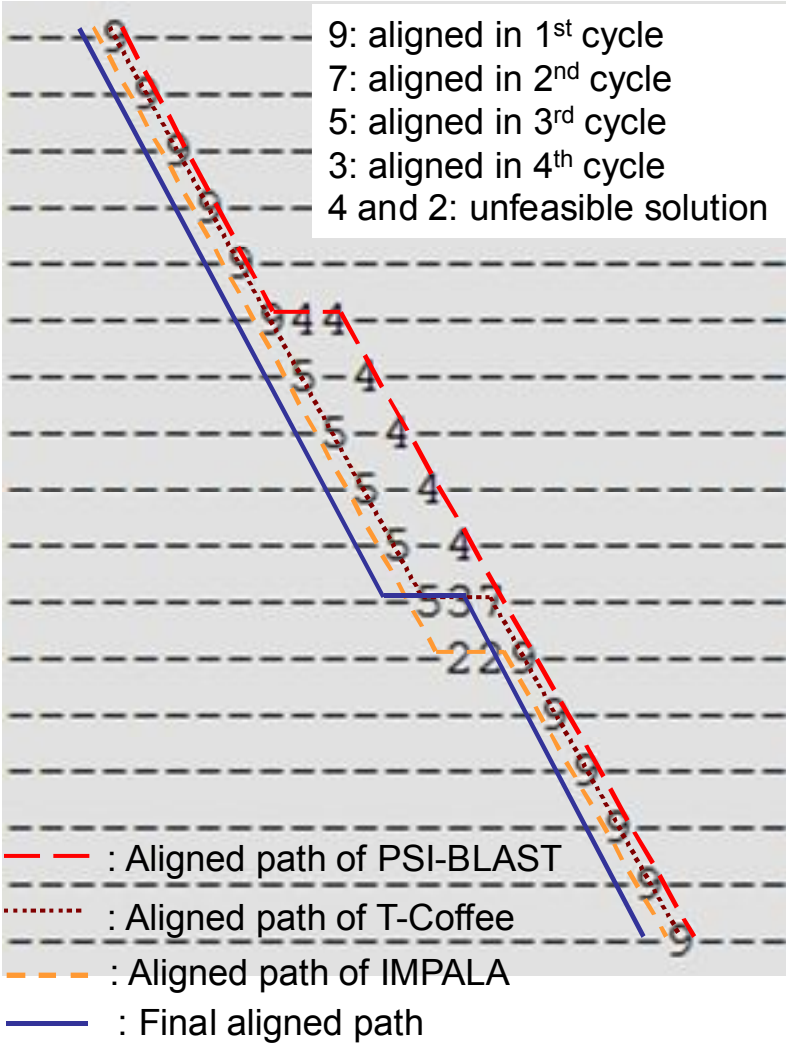
Step 3: Take the positions with the highest score as the aligned points to build the final target-template alignment. (e.g., the highest scoring is 9 for the 1st cycle in (b))

Step 4: Identify the unfeasible positions. (4 and 2 in (b))

Step 5: Change the scores of unfeasible positions and the aligned points to 0.

Step 6: Repeatedly Steps 3 and 5 until all entries are 0.

Step 7: Output the path with the aligned points as the target-template alignment



(Chen *et al.*, 2006

Materials and Methods

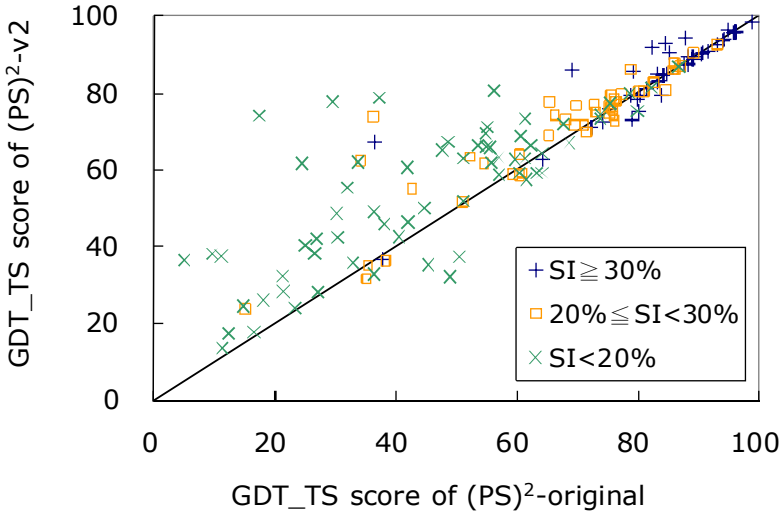
- **Performance evaluation**
  - Compare the 154 TBM targets to evaluate the performance with the other groups in CASP8.

- **GDT_TS score**

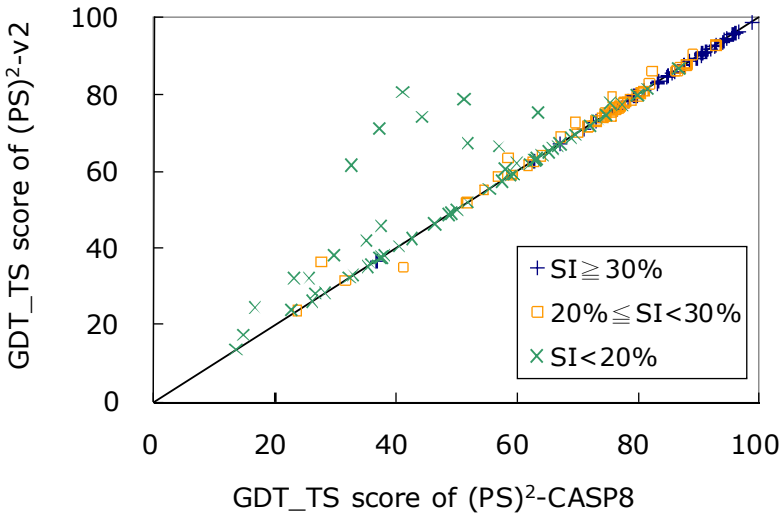
$$GDT\_TS = 100 \frac{\sum_{d} \frac{GDT_d}{N}}{4} (\%) \quad d \in \{1, 2, 4, 8\}$$

where *N* is the total number residues of a target (native structure), GDT_d is the number of aligned residues whose C α -atom distance between the target and predicted model is less than *d* Å after superposition of the two structures; and *d* is 1, 2, 4, or 8 Å.

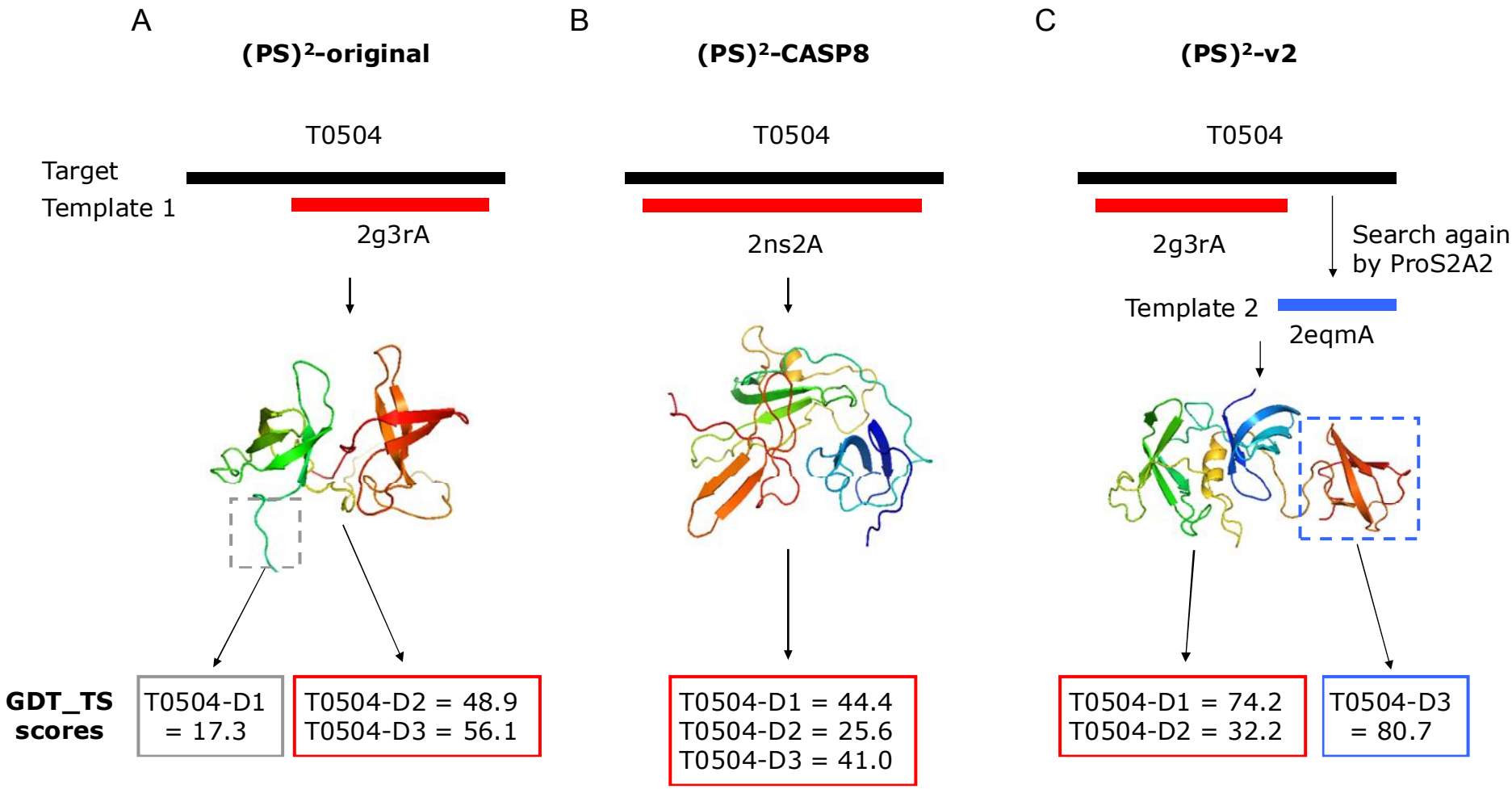
A



B



Comparison the (PS)²-v2 server with (A) (PS)²-original and (B) (PS)²-CASP8 servers on 154 TBM targets in CASP8.

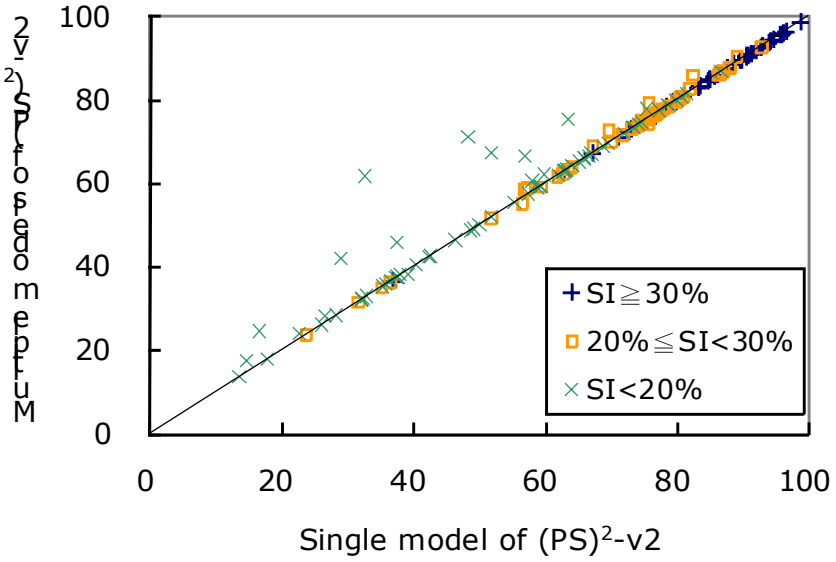


Comparison the (PS)²-v2 server with (PS)²-original and (PS)²-CASP8 servers on the target T0504 in CASP8.

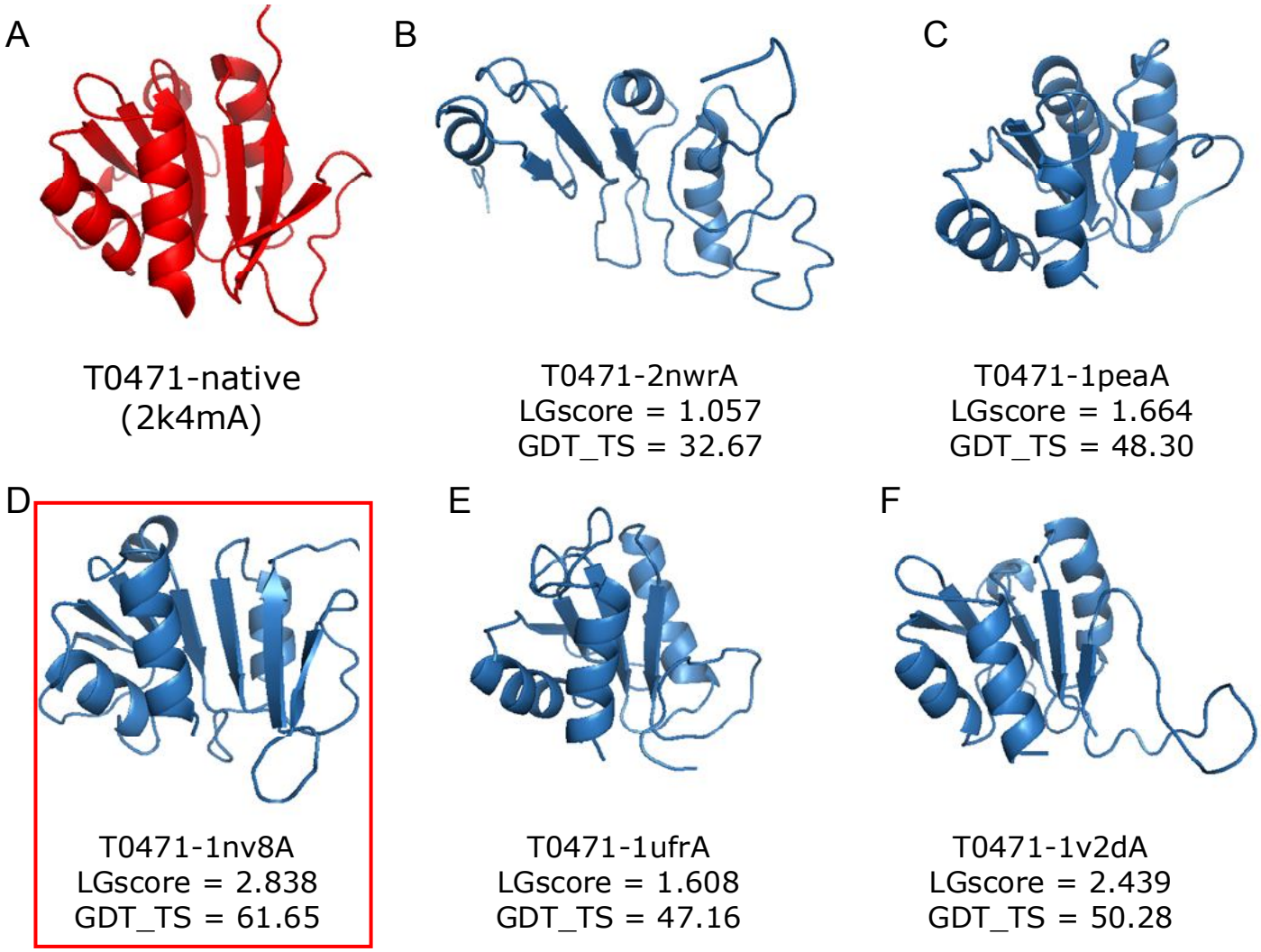
Multiple models and model selection

$(PS)^2$-v2 results for using single-model and multiple-model strategies on 154 targets in CASP8 based on GDT_TS scores

	Number of targets	Number of targets with the same GDT_TS	Number of targets improving by multiple models	Number of targets decreasing by multiple models	Sum of improving GDT_TS by multiple models	*p*-value
SI ≥ 30%	40	39	0	1	-0.4	0.3235
30%>SI≥20%	47	36	9	2	16.3	0.0231
SI <20%	67	52	14	1	129.4	0.0045
Total	154	127	23	4	145.3	0.0018



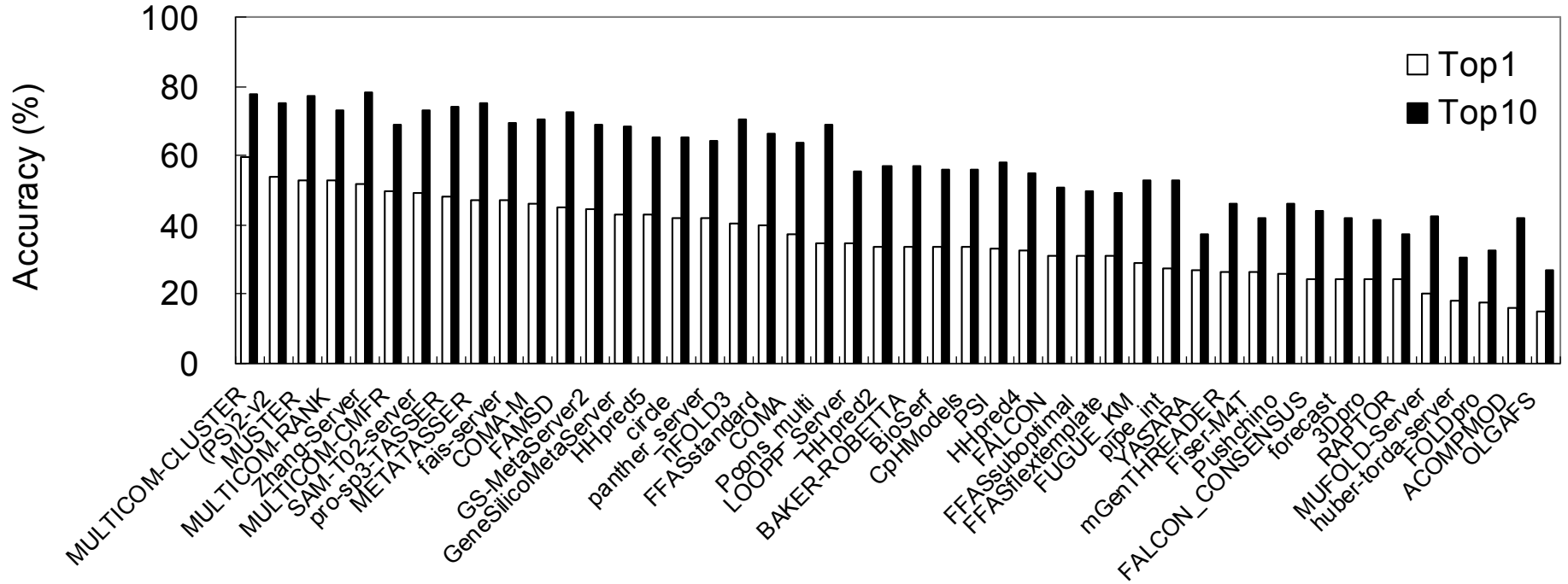
Multiple models and model selection



$(PS)^2$-v2 models the target T0471 in CASP8 using multiple models.

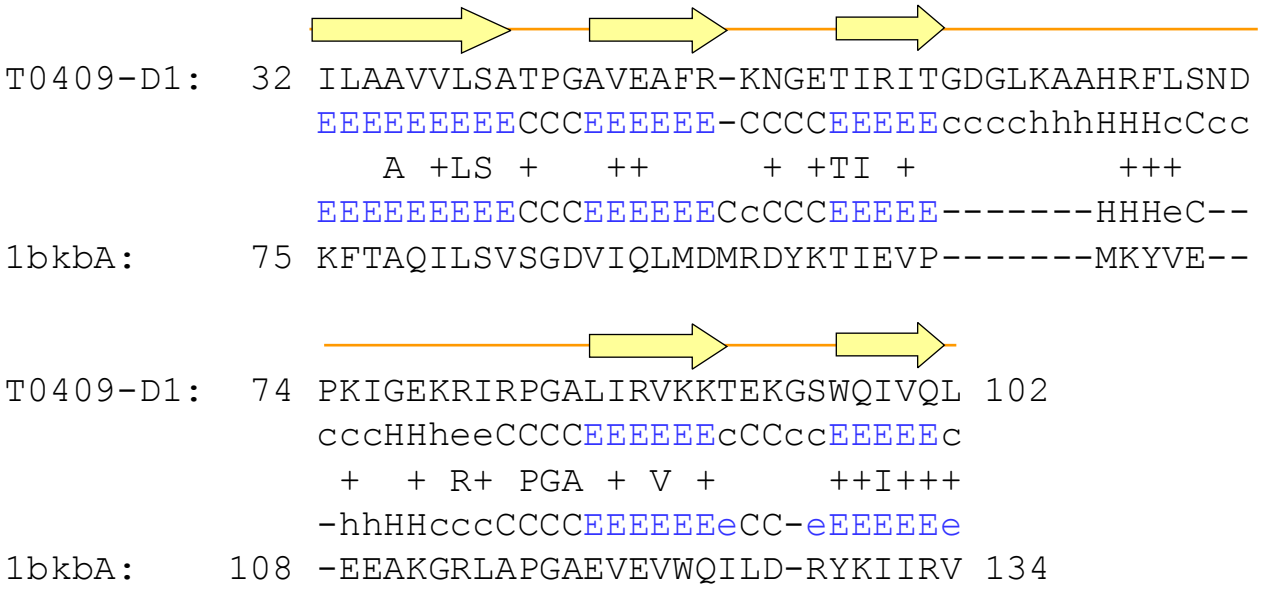
Comparing $(PS)^2$-v2 with 71 automatic servers on 154 targets in CASP8

Rank	Servers	Sum of GDT_TS score
1	Zhang-Server	10870.7
2	RAPTOR	10584.5
3	pro-sp3-TASSER, Phyre_de_novo	10469.3 ~ 10452.9
5	BAKER-ROBETTA, **$(PS)^2$-v2**, MULTICOM-CLUSTER	10358.9, **10331.4**, 10325.8
8	METATASSER	10296.7
…	…	…
…	…	…
72	mahmood-torda-server	1355.2

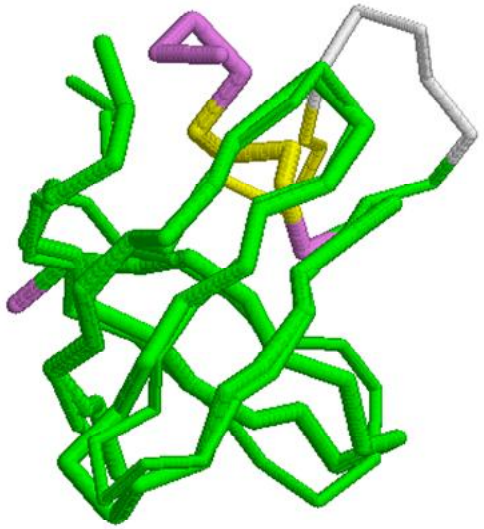


Comparison of the $(PS)^2$-v2 server with top-ranking 45 servers participating in the CASP8 competition for the template selection on 154 TBM targets.

A

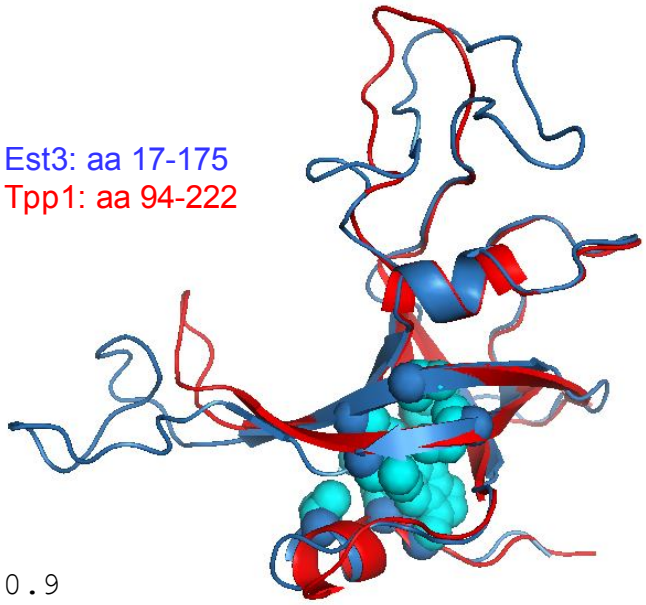


B



The alignment and predicted structure of the target T0409 using the $(PS)^2$-v2 server.

Modeling of ever shorter telomeres 3 (Est3)



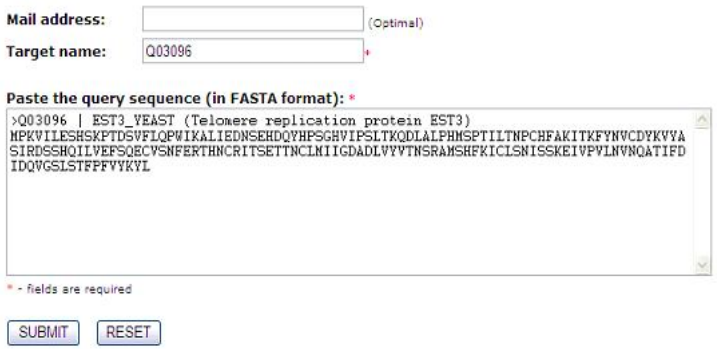

```
>Est3_YEAST <-> Tpp1 (2i46A)
 Template = 2i46A, Q_len = 181, S_len = 152, Score = 220.9
 Aligned = 87.85%, Identnties = 17.61%, SS_identities = 66.67%

Est3 :
Est3 :  17 FLQPWIKALIEDNSEHDQYHPSGHVIPSLTKQDLALPHMSPTILTNPCHFAKITKFYNVCDYKVYASIRDSSHQILVEFS 96
           EeChhHHHHccCcCCCCCeeEEEEEEEEcCCCCCccccccccCCCCCCcCCCCccccccCCCeEEEEECCCCEEEEEEC
            L+PWI+ LI    SE      +G ++  L   + A        +  P H A  T      D      + D +H +    +
           EcCccHHHHhhC-CCCCCccEEEEEEEEeCCCCC--------CCCCCC-CCCC-----CCCcEEEEECCCCEEEEEC
Tpp1 :  94 VLRPWIRELILG-SETPSSPRAGQLLEVLQDAEAA--------VAGPSH-APDT-----SDVGATLLVSDGTHSVRCLVT 158
Tpp1 :

Est3 :
Est3 :  97 QECVSNFERTHNCRITSETTNCLMIIGDADLVYVTNSRAMSHFKICLSNISSKEIVPVLNVNQATIFDIDQVGSLSTFP 175
           HHHHHhhhhhCCCeeeCCCCCEEEEEEEEEeEEEECccccccccceeeECccCcCeEEEEEEEEEEEcccCccCCCCCC
           +E +   +          T   L++ D   V+V                        L  V++ ++   +Q       P
           HHHHHccccCCCccCCCCCEEEEEEEEE-EEEEC----------cEC--CeCcEEEEEEEEEEeeeC--CCCCCC
Tpp1 : 159 REALDTSDWEEKEFGFRGTEGRLLLLQDCG-VHVQV----------AEG--GAPAEFYLQVDRFSLLPTEQ--PRLRVP 222
Tpp1 :
```

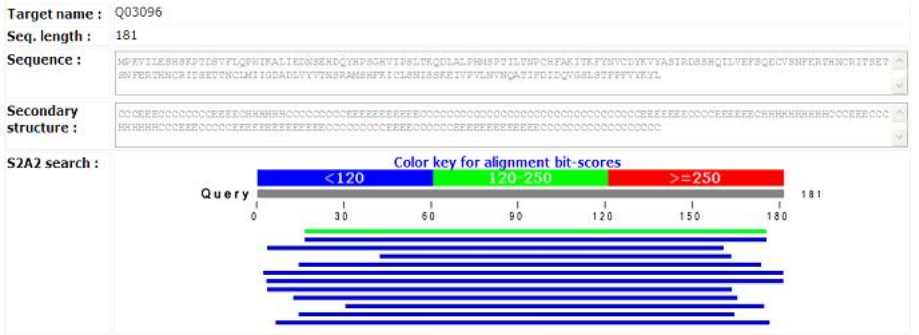

(Lee *et al.*, 2008)

A

B

http://ps2v2.life.nctu.edu.tw





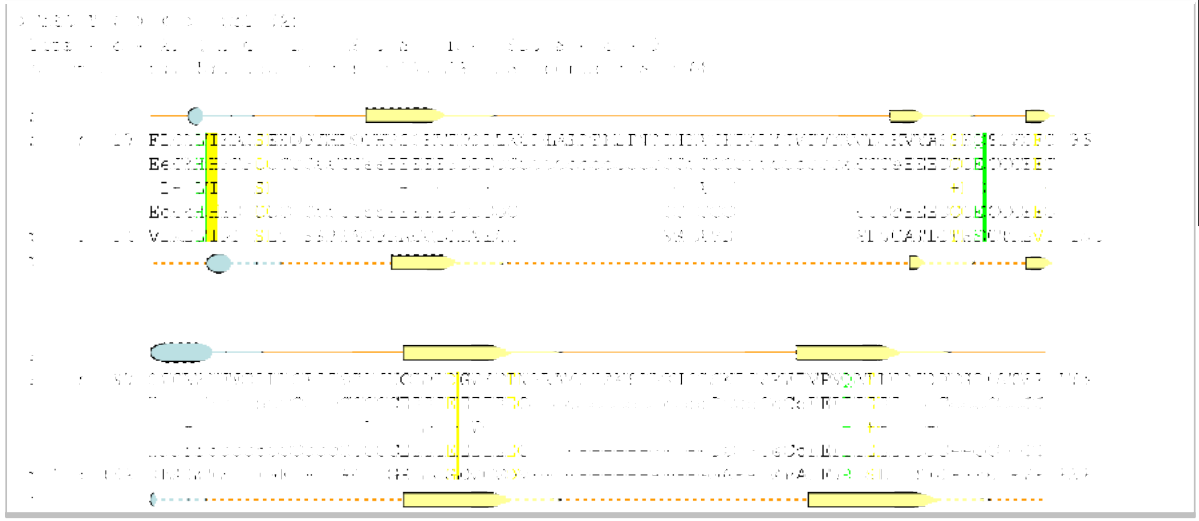
Options:

Template(s) selection: Automatic, Manual (users can select the modeling template by themself) or Use this template: xxxx_y (where xxxx is the PDB code and y is the chain)

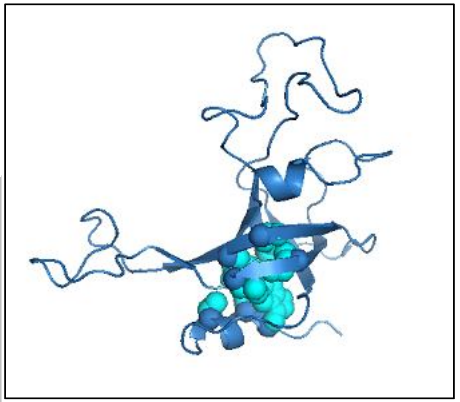
Template	Seq-len	Aligned (%)	Identity (%)	SW-score	Bit-score	E-value
2i46A	152	87.85	17.61	220.9	120.3	0.014
2i46B	134	87.85	13.84	190.9	105.5	0.096
2qf4A	170	86.74	16.67	174.2	94.8	0.38
2a22B	203	66.85	12.21	174.8	94.5	0.39
3c7fA	487	87.85	20.47	187.4	93.0	0.47

C

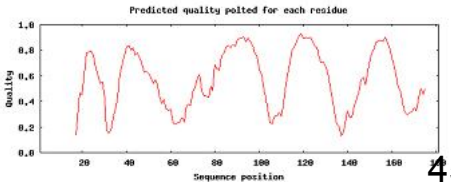
PDB Model	Template	Seq-len	Aligned (%)	Identity (%)	Bit-score	E-value	ProQ (LG,MaxSub)	ProQres	Structure
model_1	2i46A	152	87.85	17.61	120.3	0.014	1.664, 0.389		

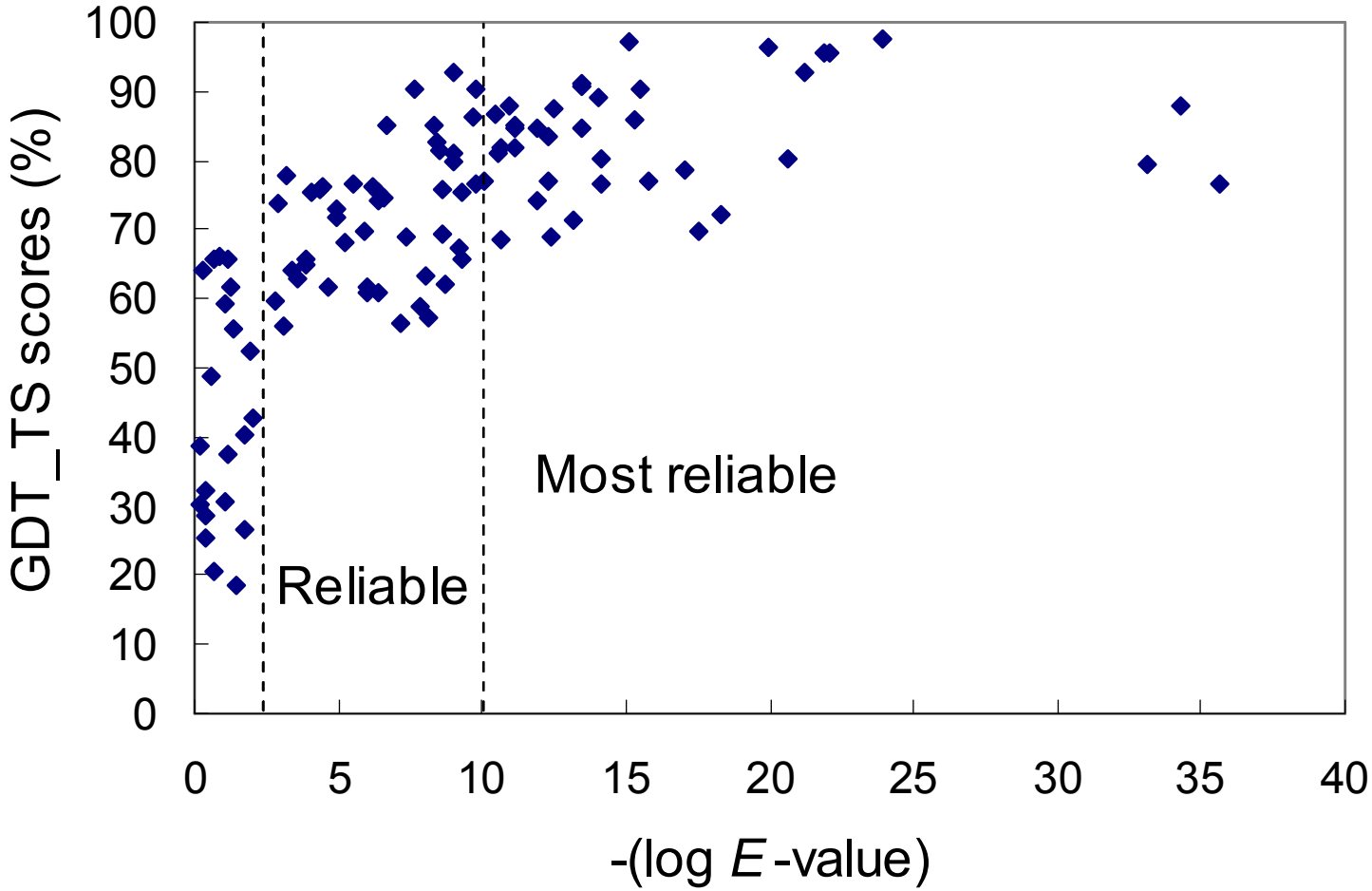


D



E





The relation between *E*-values and GDT_TS scores of $(PS)^2$-v2 for the targets in CASP8.

Conclusion

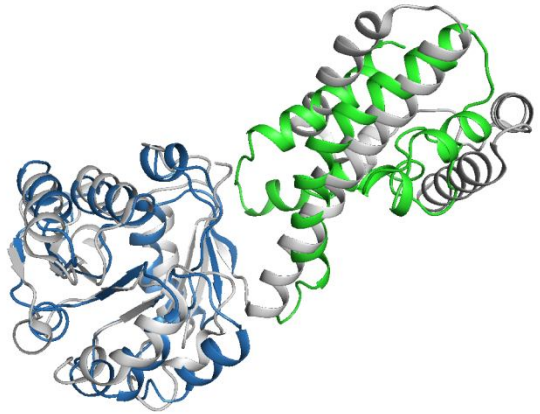
- We introduce a new way to incorporate sequence and secondary structure into a single substitution matrix (**S2A2**).
- We also develop the **ProS2A2** alignment method that both combines **S2A2**, and **PSSM** to detect the remote homologs in functional prediction and database search.
- The accuracy of ProS2A2 is better then the **sequence profile methods** and is comparable to the **structure based methods**.

Conclusion

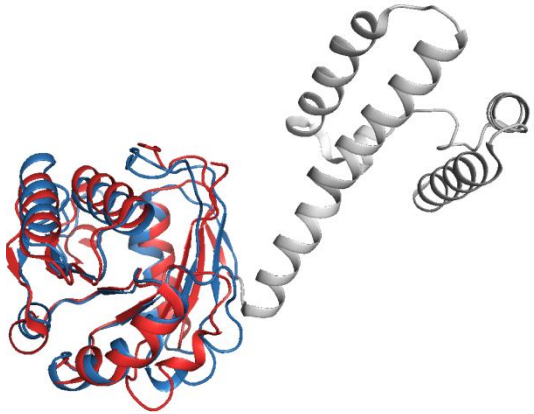
- Experimental results demonstrate that $(PS)^2$-v2 with the ProS2A2 is useful for **template selections** and **target-template alignments** by blending the amino acid and structural propensities.
- The **multiple-template** and **multiple-model** strategies are able to significantly improve the accuracies for protein structure prediction in the twilight zone.

Discussion

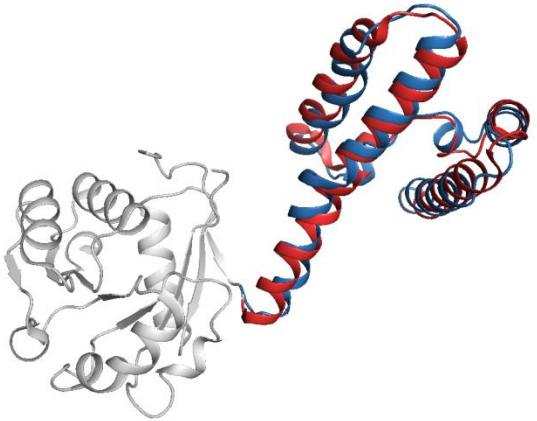
- Our poor performance was partially due to the prediction without any **long loop modeling** and **domain parsing**.



T0393-D1, **T0393-D2**
Template: 2h78A
GDT_TS = 59.3, 31.6
Rank: #50, #56



T0393-D1
Template: 2f1kA
GDT_TS = 68.1
Rank: #8



T0393-D2
Template: 2i76A
GDT_TS = 71.7
Rank: #16

Thanks for your attention!