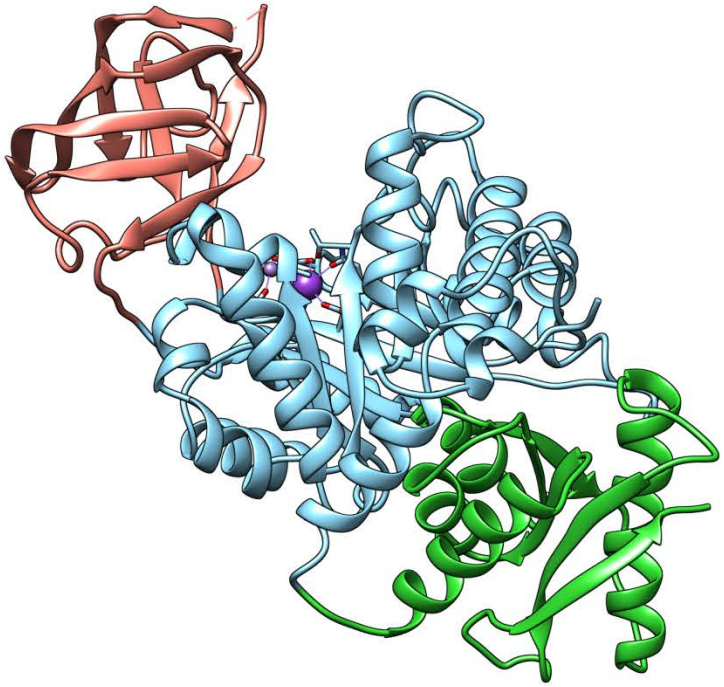
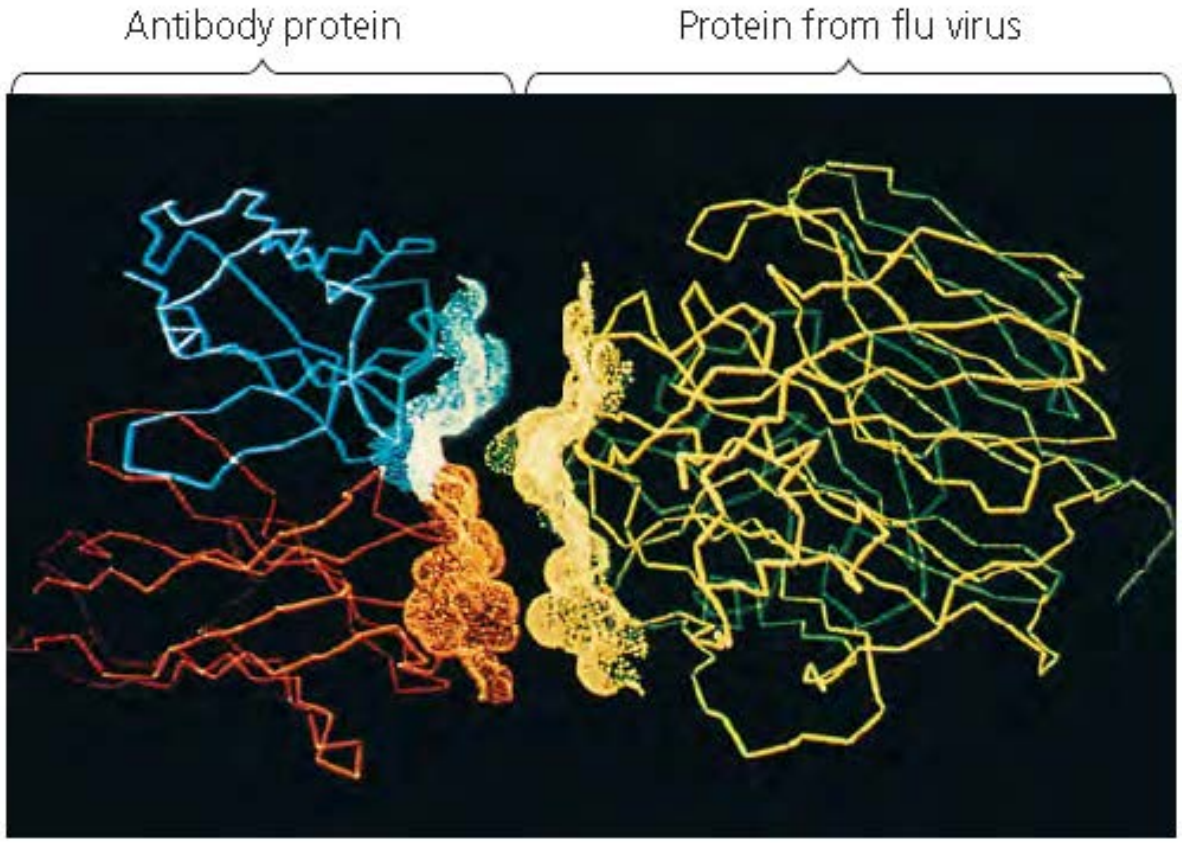


June 20, 2014

Protein domain prediction and modeling unit determination for protein structure prediction by using sequence and structure information

Kyuhong Choe
Seoul National University
Interdisciplinary program in Bioinformatics
Lab of Computational Biologyand Biomolecular Engineering

Why do we need protein structure?



Reece *el al.*, Campbell Biology, 9th edition, *PEARSON*,

Why "predict" protein structure?

Experimental methods are expensive and time-consuming.
Many proteins are hard to prepare for experimental structure determination.

Available protein sequences: 38,633,935
Available protein structures: 100,843

http://www.ncbi.nlm.nih.gov/refseq/	May 12, 2014	(RefSeq)
http://www.wwpdb.org	May 13, 2014	(wwpdb)

Design+

Structure Prediction Methods

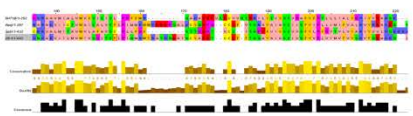
Ab initio modeling:
Based on physical principles

Template-based modeling (TBM):
By selecting templates from the database of known experimental structures
→ More accurate

Template-based Modeling (TBM)

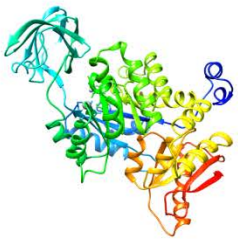
MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYPPQG
HGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQGGGTHSQWNKPSKPKTNMKHMA
VVGGLGGYMLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQN
NITIKQHTVTTTTKGENFTETDVKMMERVVEQMCITQYERESQAYYQRGSSMVI
ILLISFLIFLIVG

Sequence



Sequence Alignment

Fold Recognition



3D model building

GalaxyTBM (Seok lab @SNU)

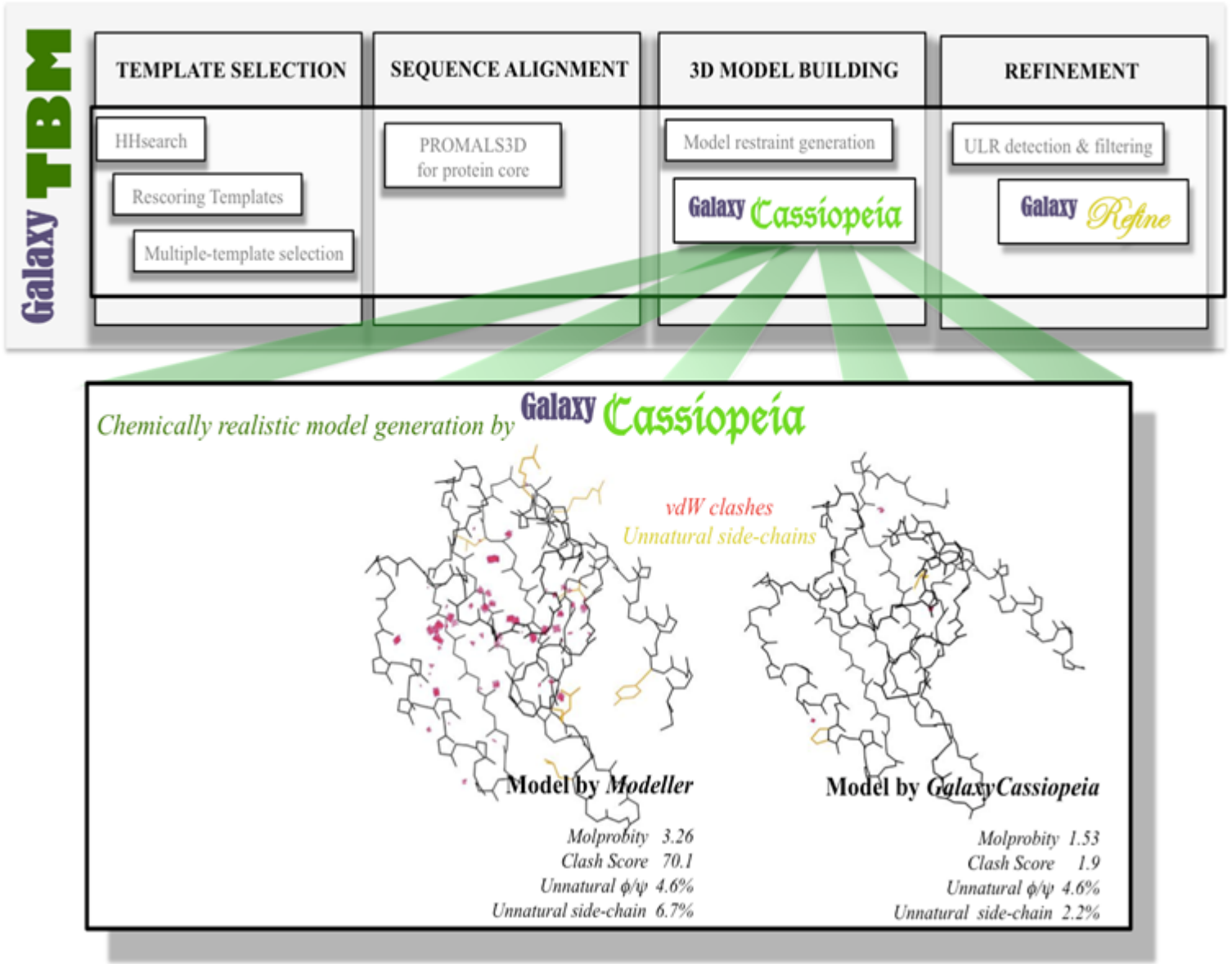
Performance comparison for single-domain targets

Target	GalaxyTBM			MODELLER[1]			SWISS-MODEL[2]		
	GDT-TS (%)	GDC-SC(%)	Mol-Probity	GDT-TS (%)	GDC-SC (%)	Mol-Probity	GDT-TS (%)	GDC-SC(%)	Mol-Probity
T0516	74.34	**35.11**	**1.72**	**74.45**	29.59	3.06	73.90	31.03	2.94
T0591	**76.05**	**36.16**	**2.39**	75.27	31.83	3.41	73.36	28.30	3.08
T0597	**76.18**	**40.64**	**2.28**	72.77	34.69	3.37	73.33	28.30	3.17
T0609[3]	**68.21**	**27.89**	**2.35**	67.76	26.62	3.81	-	-	-
T0641	72.37	**31.84**	**2.62**	71.10	29.18	3.66	**72.46**	32.34	3.05
T0650	**85.40**	**50.39**	**1.87**	85.33	44.03	3.17	56.56	45.15	3.01
T0652	**94.58**	**47.04**	**1.67**	92.77	42.44	2.93	87.95	32.99	2.54
T0658	**80.56**	**45.06**	**2.57**	80.43	39.99	3.55	16.41	10.46	3.40
T0682[3]	**79.81**	**42.30**	**2.17**	78.82	34.71	3.11	-	-	-
T0749	**91.20**	**57.45**	**2.08**	89.86	53.14	3.08	90.11	55.89	2.69
Average	**79.87**	**41.39**	**2.17**	78.86	36.62	3.31	-	-	-

[1] MODELLER version 9.11 was used with the same multiple sequence alignment as GalaxyTBM.
[2] SWISS-MODEL web server was used with the same template list as GalaxyTBM.
[3] Targets for which SWISS-MODEL web server failed to generate models.

GDT-TS, GDC-SC : the bigger, the better MolProbity : the smaller, the better

Shin, W., Lee, G. R., Heo, L., Lee, H. & Seok, C. Prediction of Protein Structure and Interaction by GALAXY Protein Modeling Programs. *Bio Des.* **2,** 1–11 (2014).

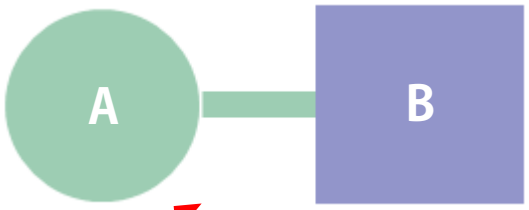


Shin, W., Lee, G. R., Heo, L., Lee, H. & Seok, C. Prediction of Protein Structure and Interaction by GALAXY Protein Modeling Programs. *Bio Des.* **2,** 1–11 (2014).

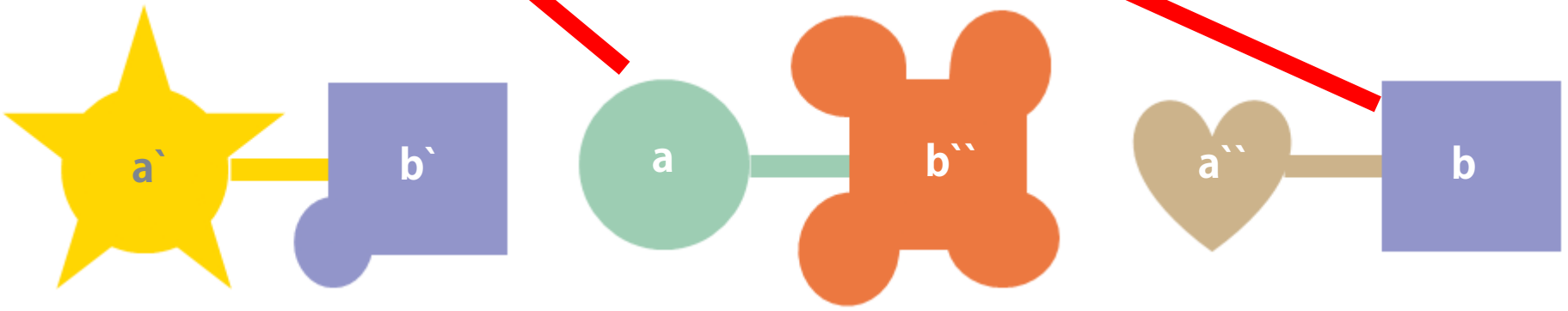
Multi-domain protein targets: CASE I

Design+

Target protein (structure to be predicted)

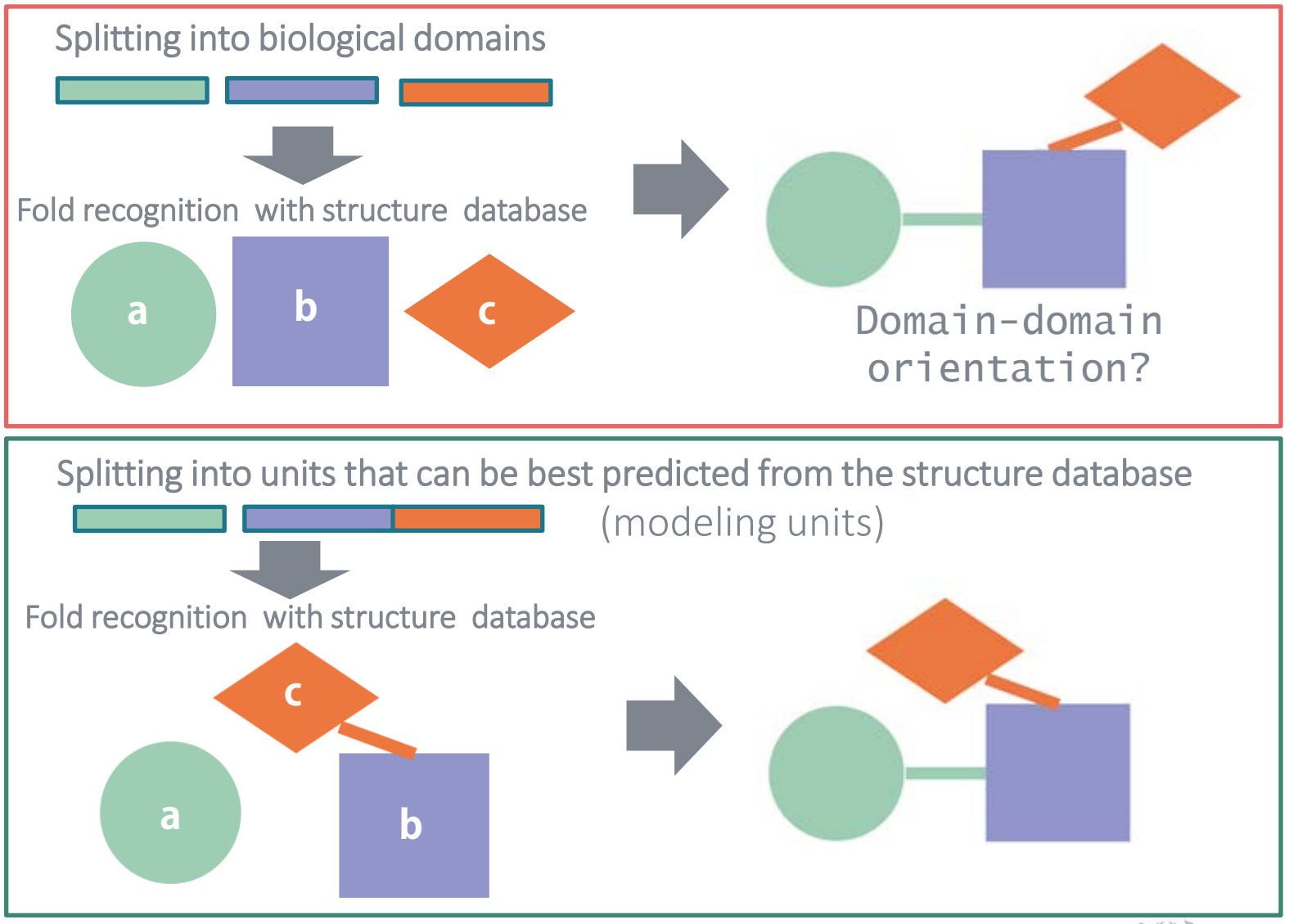


Structure database



Multi-domain protein targets: CASE II

Design+



SCIENCE
Seoul National University

GalaxyDom (ver.1)

Design+

Template-based protein domain prediction

Use Hhsearch(sequence alignment with predicted secondary structure information) for fold recognition*

→ 30 highest-ranking proteins

→ global & local alignment

→ find regions with less gaps

by introducing "chunk score" C_i

(estimates the probability of the i-th residue to be a part of a chunk or a domain)

$$C_i = \langle G_j L_j \rangle_{i-7,i+7}$$

i: *residue number*

G_j: *a score from global alignment result*

L_j: *a score from local alignment result*

SCIENCE Seoul National University

* Söding, J. Protein homology detection by HMM–HMM comparison. *Bioinformatics* **21,** 951–60 (2005).

GalaxyDom (ver.1)

Design+

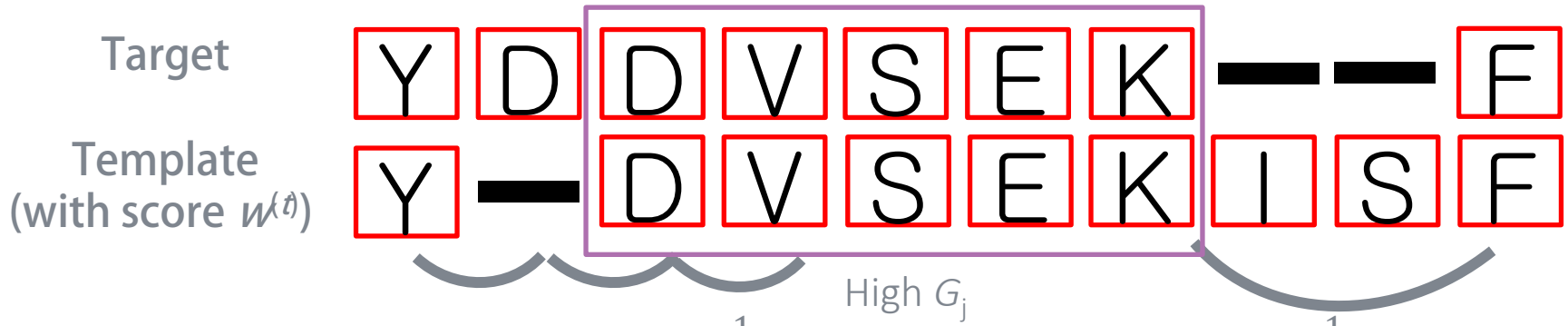
Global alignment score

$$G_j = \sum_{t}^{N_t} w^{(t)} \cdot g_{j,j+1}^{(t)}$$

N_t : *number of templates*

$w^{(t)}$: *weight factor of the t-th template (calculated from alignment score)*

$g_{j,j+1}^{(t)}$: *alignment score of target and template for (j, j+1) residue pair*



g = 0 0 $\frac{1}{1+\left(\frac{1}{30}\right)^2}$ ..... $\frac{1}{1+\left(\frac{3}{30}\right)^2}$

d=1 d=3

$$g_{j,j+1}^{(t)} = \begin{cases} 0 \\ \dfrac{1}{1+\left(\dfrac{d_{j,j+1}^{(t)}}{30}\right)^2} \end{cases}$$

$d_{j,j+1}^{(t)}$: *distance of j and (j+1) in the alignment*

GalaxyDom (ver.1)

Local Alignment score

Design+

$$L_j = \sum_{t}^{N_t} \sum_{a}^{N_{LA}^{(t)}} \frac{1}{1 + w^{(t)} \cdot w_a^{(t)} \cdot l_{j,a}^{(t)}}$$

N_t : *number of templates*

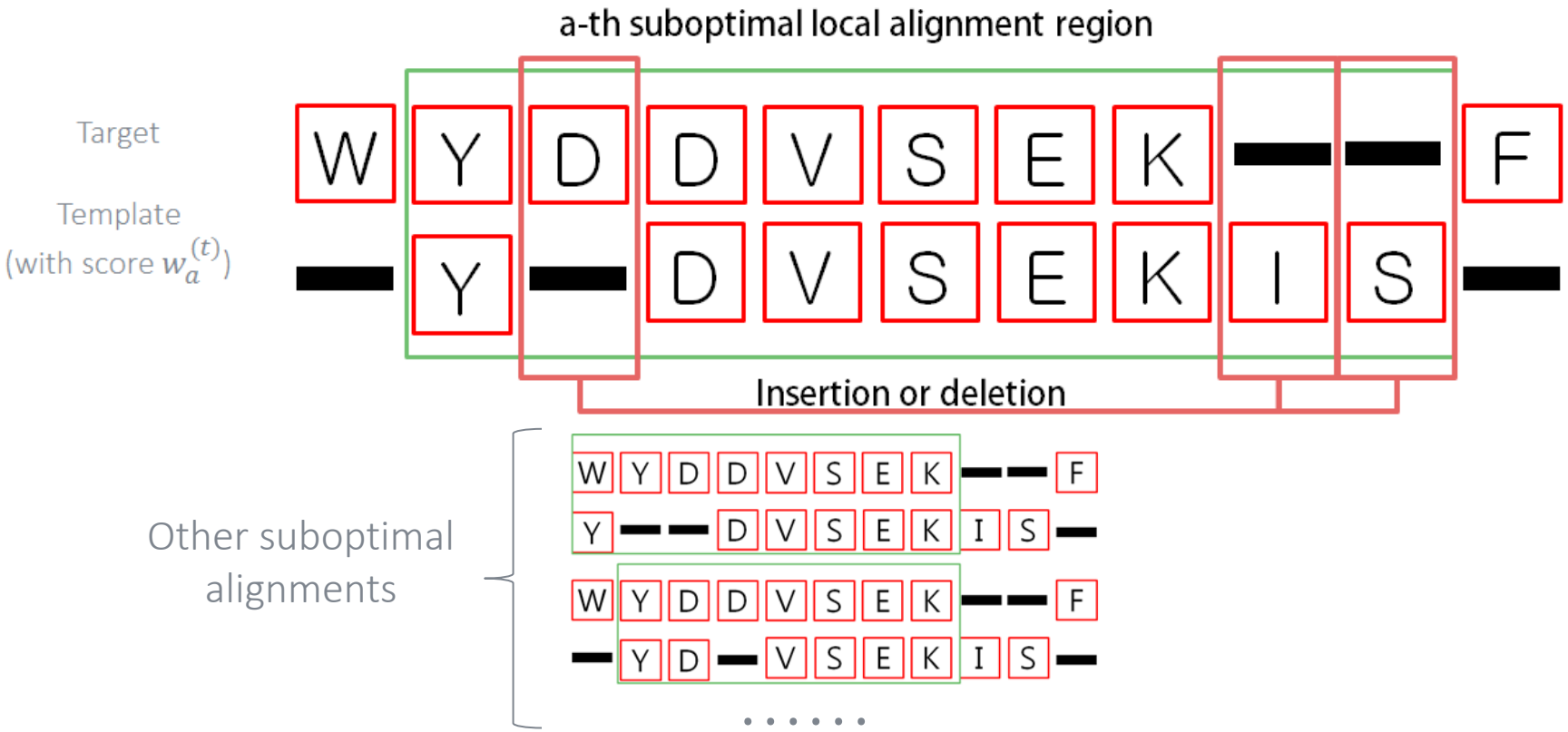
$N_{LA}^{(t)}$: *number of alternative local alignments*

$w^{(t)}$: *weight factor of the t-th template*

$w_a^{(t)}$: *weight factor of the a-th suboptimal alignment*

$l_{j,a}^{(t)}$: *mis-alignment score*

(insertion or deletion = 1, aligned = 0)



GalaxyDom (ver.1)

Design+

Local Alignment score(Cont'd)

$$L_j = \sum_{t}^{N_t} \sum_{a}^{N_{LA}^{(t)}} \frac{1}{1 + w^{(t)} \cdot w_a^{(t)} \cdot l_{j,a}^{(t)}}$$

N_t : *number of templates*

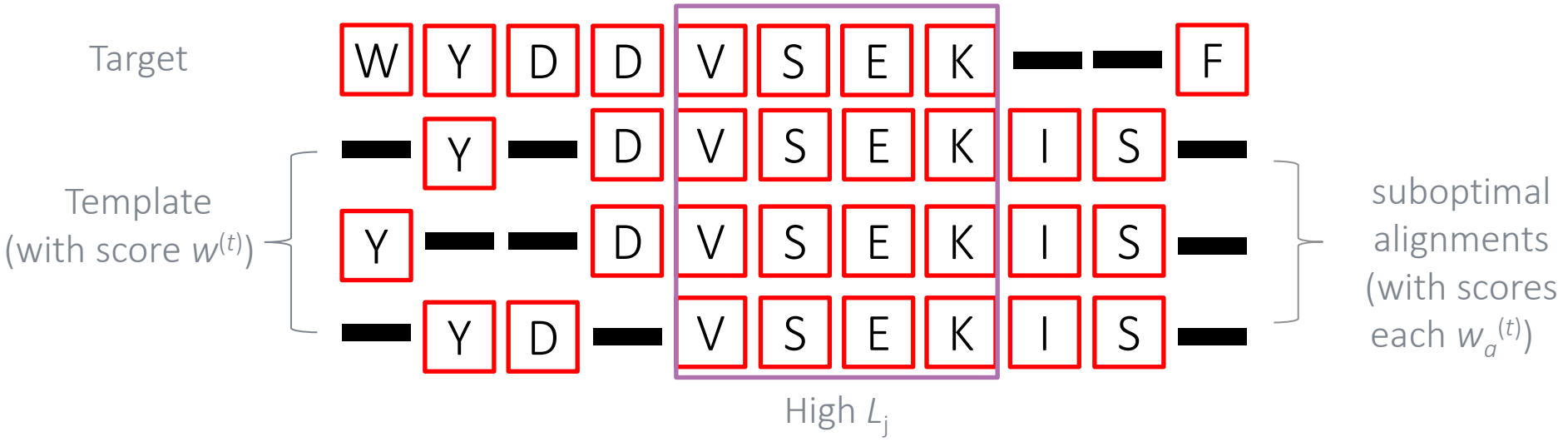
$N_{LA}^{(t)}$: *number of alternative local alignments*

$w^{(t)}$: *weight factor of the t-th template*

$w_a^{(t)}$: *weight factor of the a-th suboptimal alignment*

$l_{j,a}^{(t)}$: *mis-alignment score*

(insertion or deletion = 1, aligned = 0)



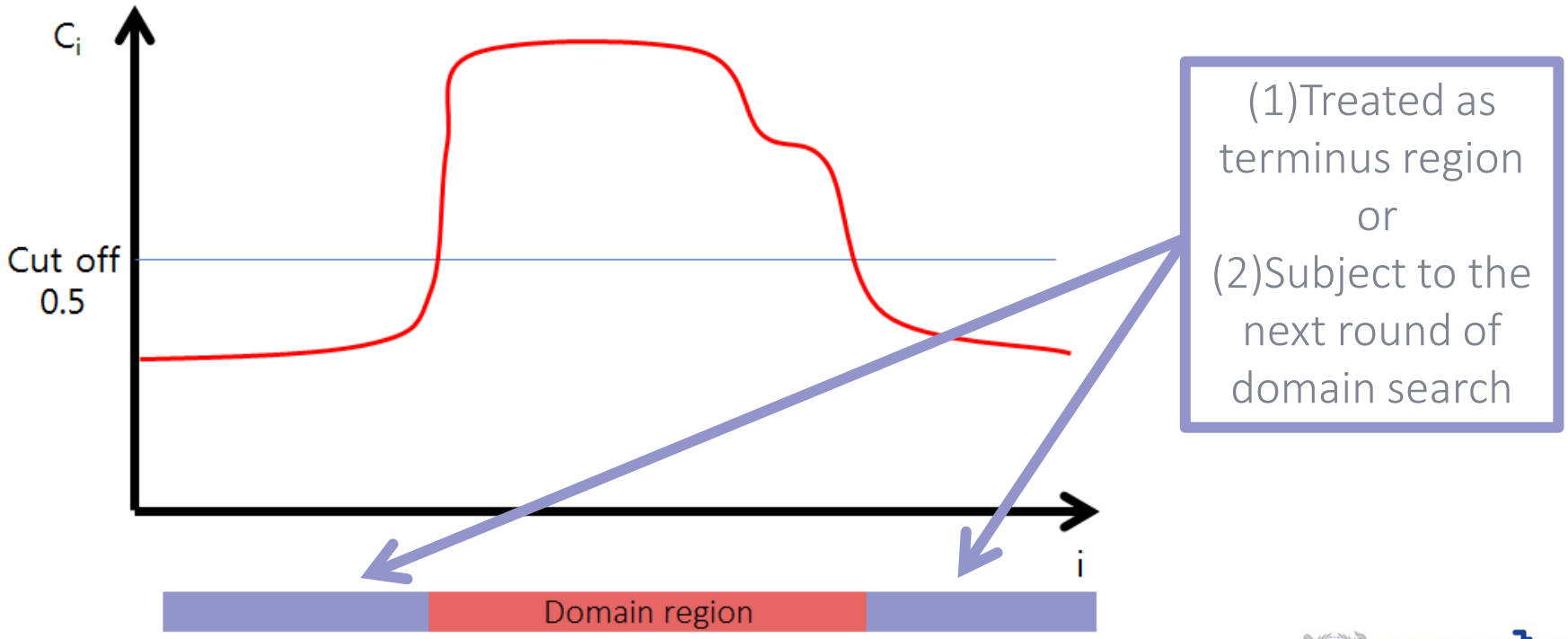
SCIENCE
Seoul National University

GalaxyDom (ver.1)

Design+

Combine the Local alignment score
and the Global alignment score for Chunk score

$$C_i = \left\langle G_j L_j \right\rangle_{i-7,i+7}$$

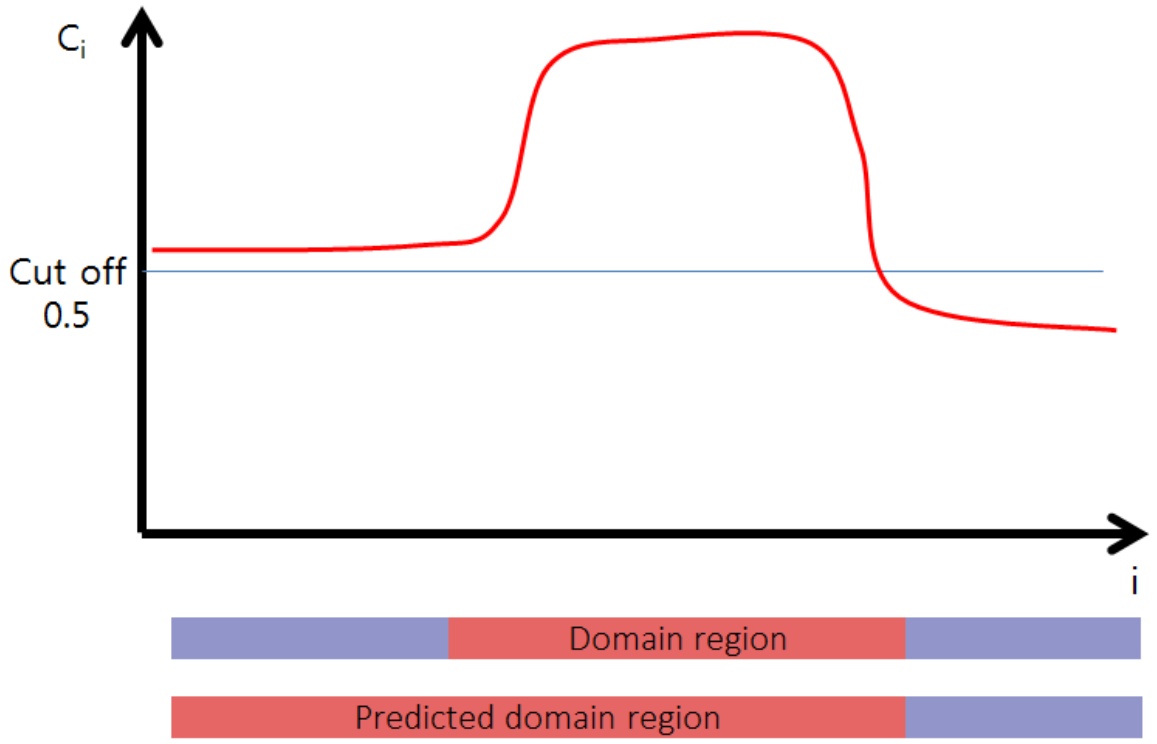


GalaxyDom (ver.2)

Design+

Problem with GalaxyDom (ver.1)

1. Under-splitting of domains

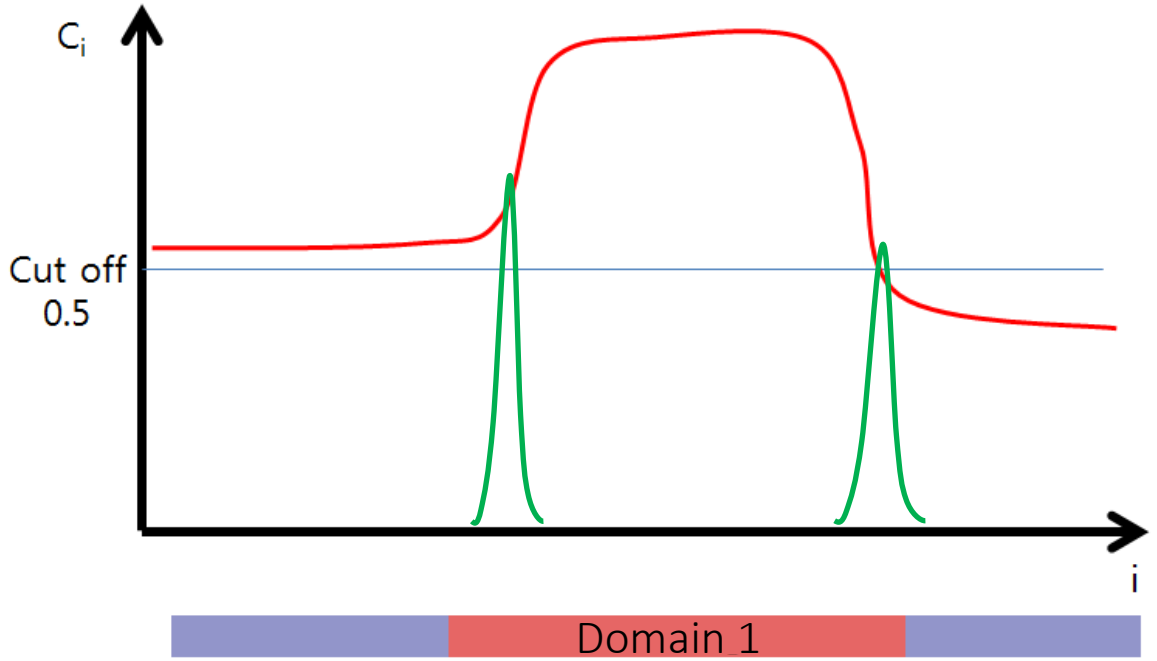


GalaxyDom (ver.2)

Design+

Problem with GalaxyDom (ver.1)

1. Under-splitting of domains



→ Introduce new method that use **derivative of chunk score** to determine domain boundaries

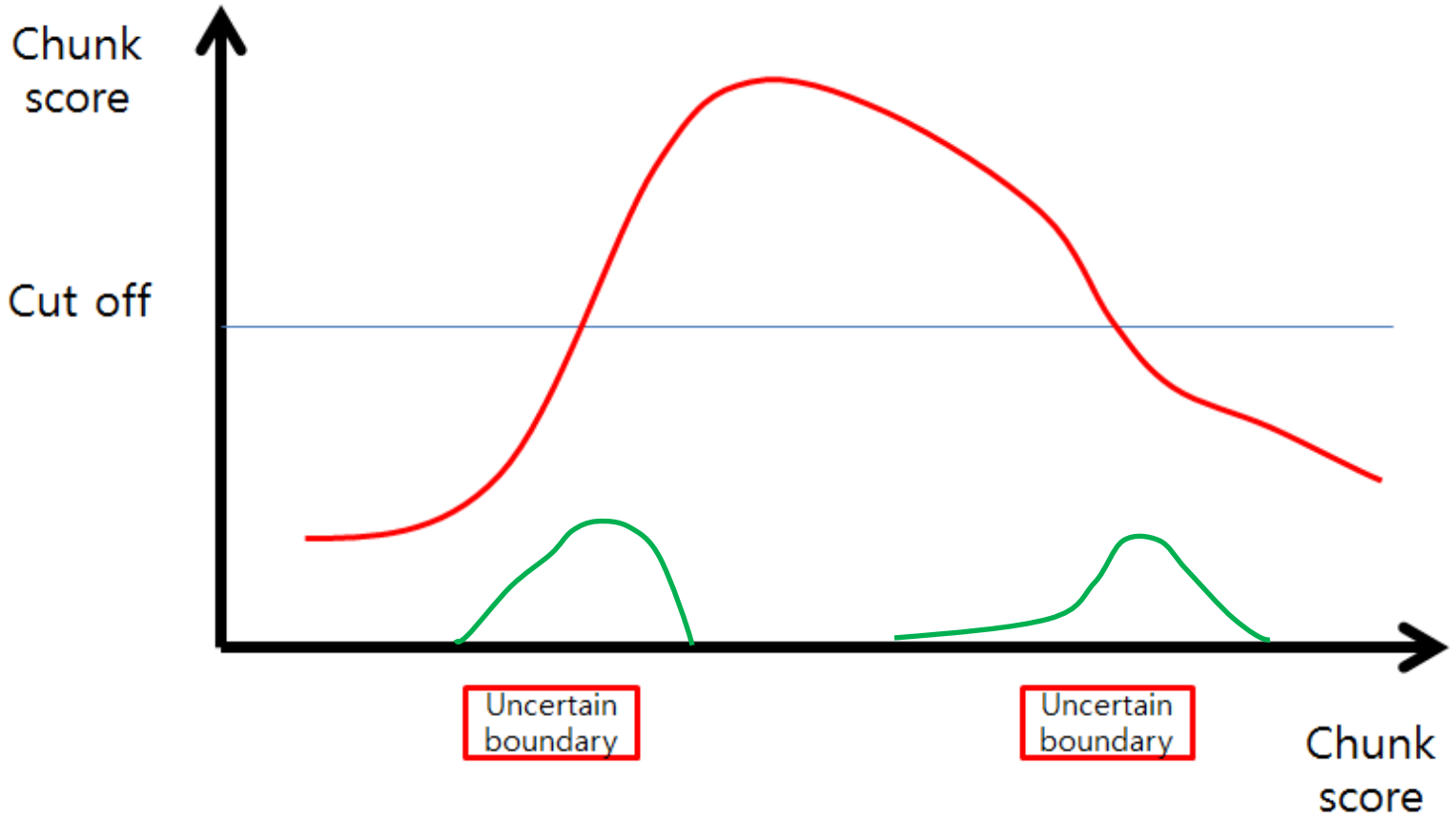
SCIENCE
Seoul National University

GalaxyDom (ver.2)

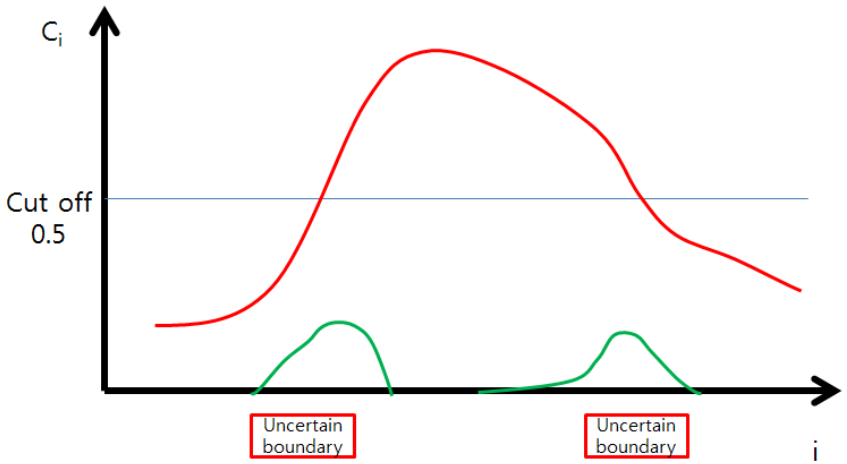
Design+

Problem with GalaxyDom (modified with gradient ver.1)

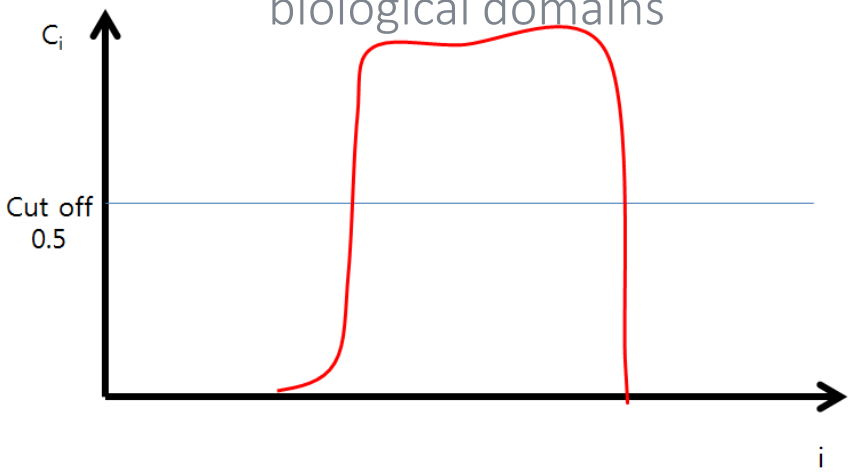
2. Uncertain domain boundary

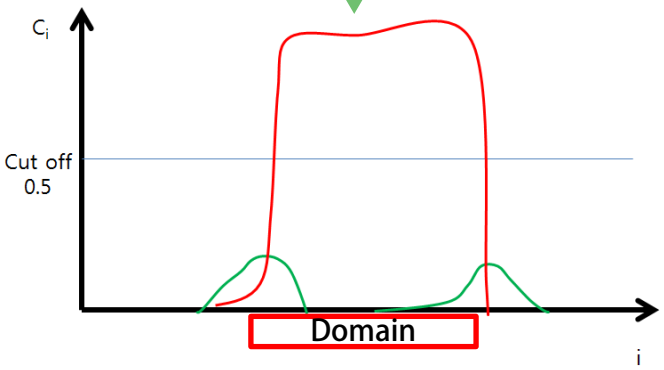


Derivative of Chunk score
From templates of the PDB database



Chunk score
From templates of the SCOP database of biological domains





→ Combine derivative of chunk score from PDB with biological domain chunk score from the SCOP database*

* Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247,** 536–540 (1995).

GalaxyDom result

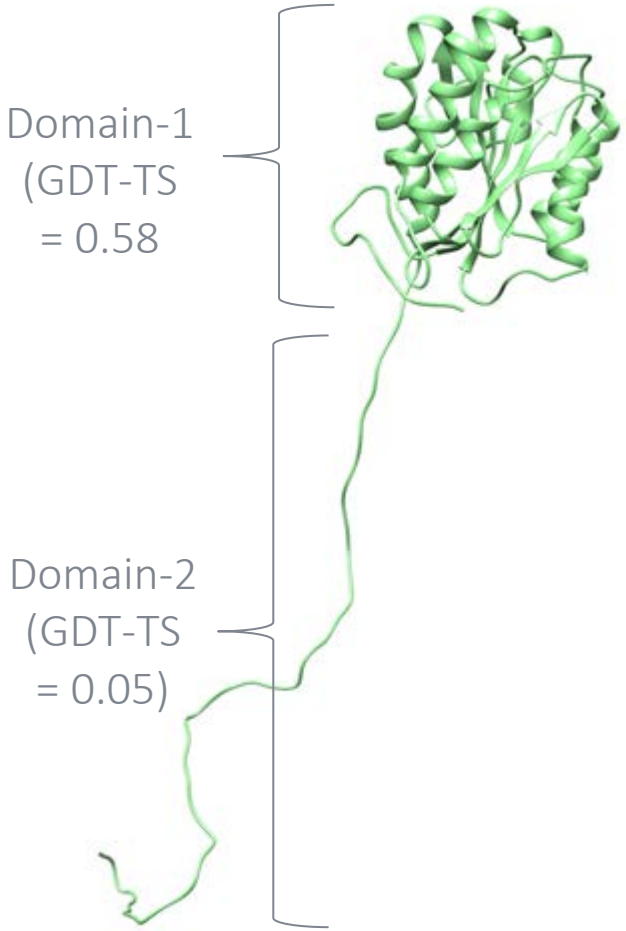
Training set: CASP9 target (# of targets = 116, # of multi-domain targets = 21)
CASP: Critical Assessment of protein Structure Experiment
a community-wide blind prediction experiment

Test set: CASP 10 target (# of targets = 97, # of multi-domain targets = 22)

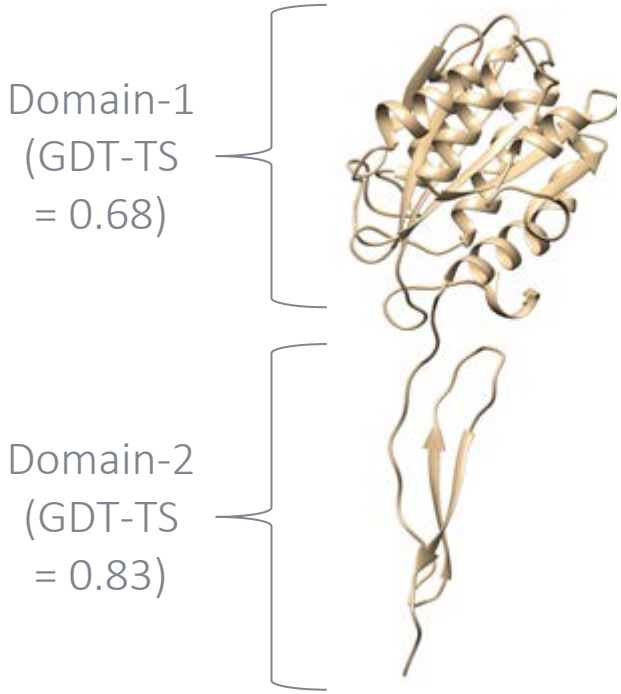
	GalaxyDom (ver.1)	GalaxyDom (ver.2)
Successful Prediction	70/97(72.9%)	76/97(79.7%)

(success: domain boundaries predicted
within CASP domain definition ± 10 residues)

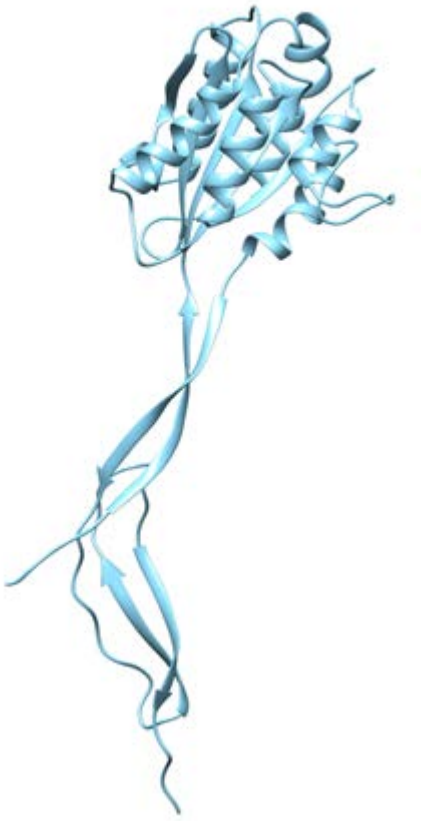
GalaxyDom result: a Successful Example



Model generated
without domain splitting



Model generated
after domain prediction
with GalaxyDom



X-ray diffraction structure
from PDB data

Acknowledgement

Seoul National University

Lab of Computational Biology and Biomolecular engineering

June 20, 2014

Prof. Chaok Seok

Dr. Junsu Ko (Senior researcher of Theragen Bio Institute)

Lim Heo

Thank you for your attention!