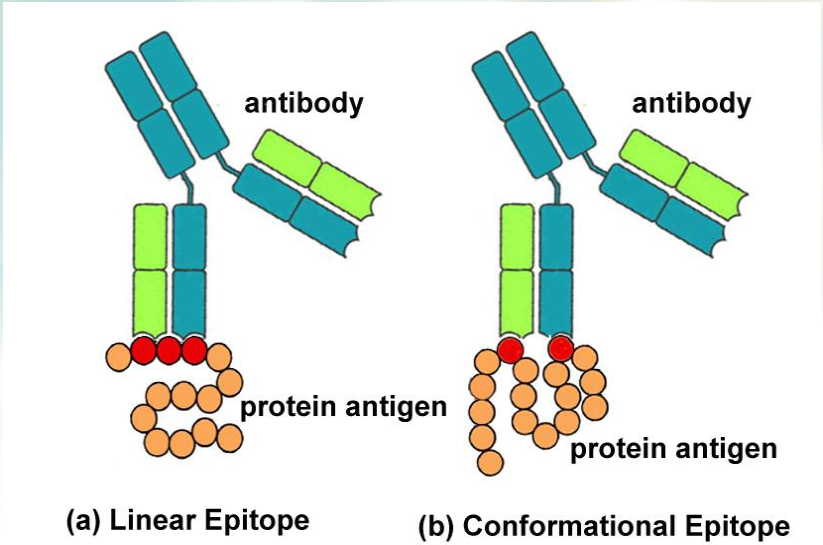
Vaccine Investigation and Online Information Network
(ICoVax 2012 2012/10/13)

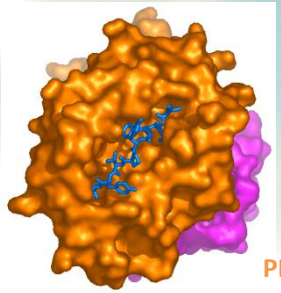
Prediction of Conformational Epitopes by Knowledge-based Energy Function and Geometrical Neighbouring Residue Contents

Ying-Tsang Lo[1], Tun-Wen Pai[1,2*], Wei-Kuo Wu[1], Hao-Teng Chang[3,4*]

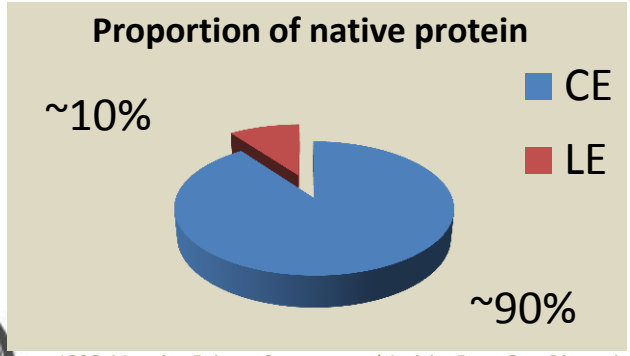
[1]Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung, Taiwan
[2]Center of Excellence for Marine Bioenvironment and Biotechnology, National Taiwan Ocean University, Keelung, Taiwan
[3]Graduate Institute of Molecular Systems Biomedicine, China Medical University, Taichung, Taiwan
[4]China Medical University Hospital, Taichung, Taiwan

Knowledge of Epitopes



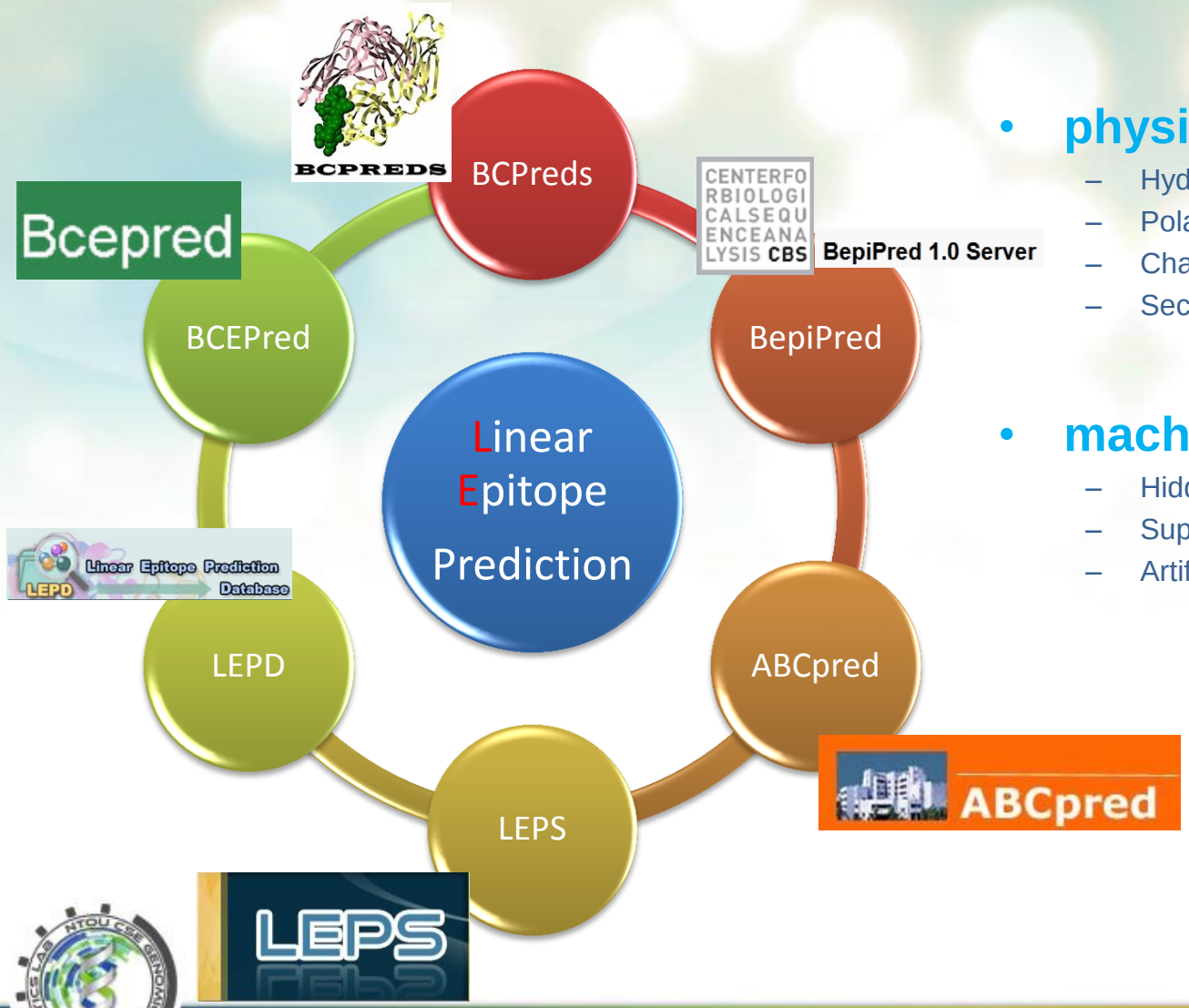


- B-cell epitopes can interact with an antigen to elicit either cellular or humoral immune response.
- In general, epitopes can be categorized into 2 types.
  - Linear (or continuous) Types
  - Conformational (or discontinuous) Types



1996_Mapping Epitope Structure and Activity From One-Dimensional Prediction to Four-Dimensional Description of Antigenic Specificity

Linear Epitope Prediction System



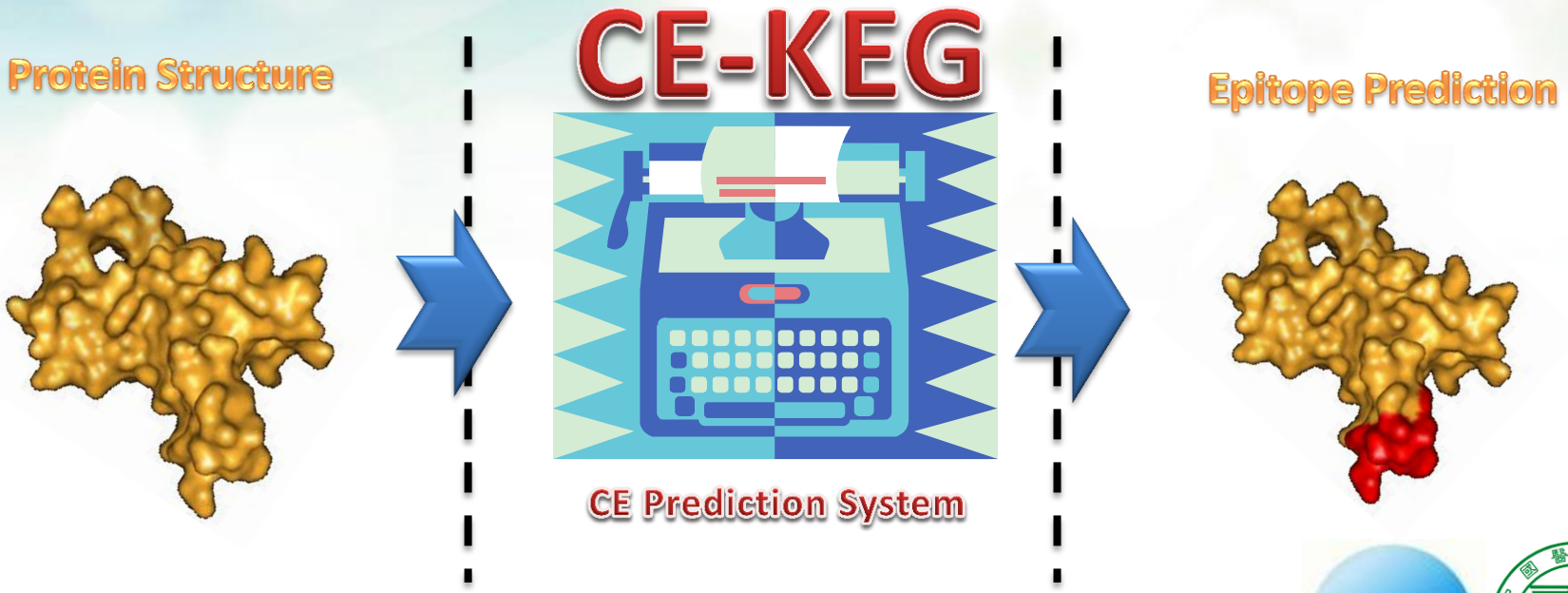
- **physical-chemical propensity**
  - Hydrophilicity
  - Polar
  - Charge
  - Secondary structure
- **machine learning algorithms**
  - Hidden Markov Model (HMM)
  - Support Vector Machine (SVM)
  - Artificial Neural Network (ANN)

Conformational Epitope Prediction System

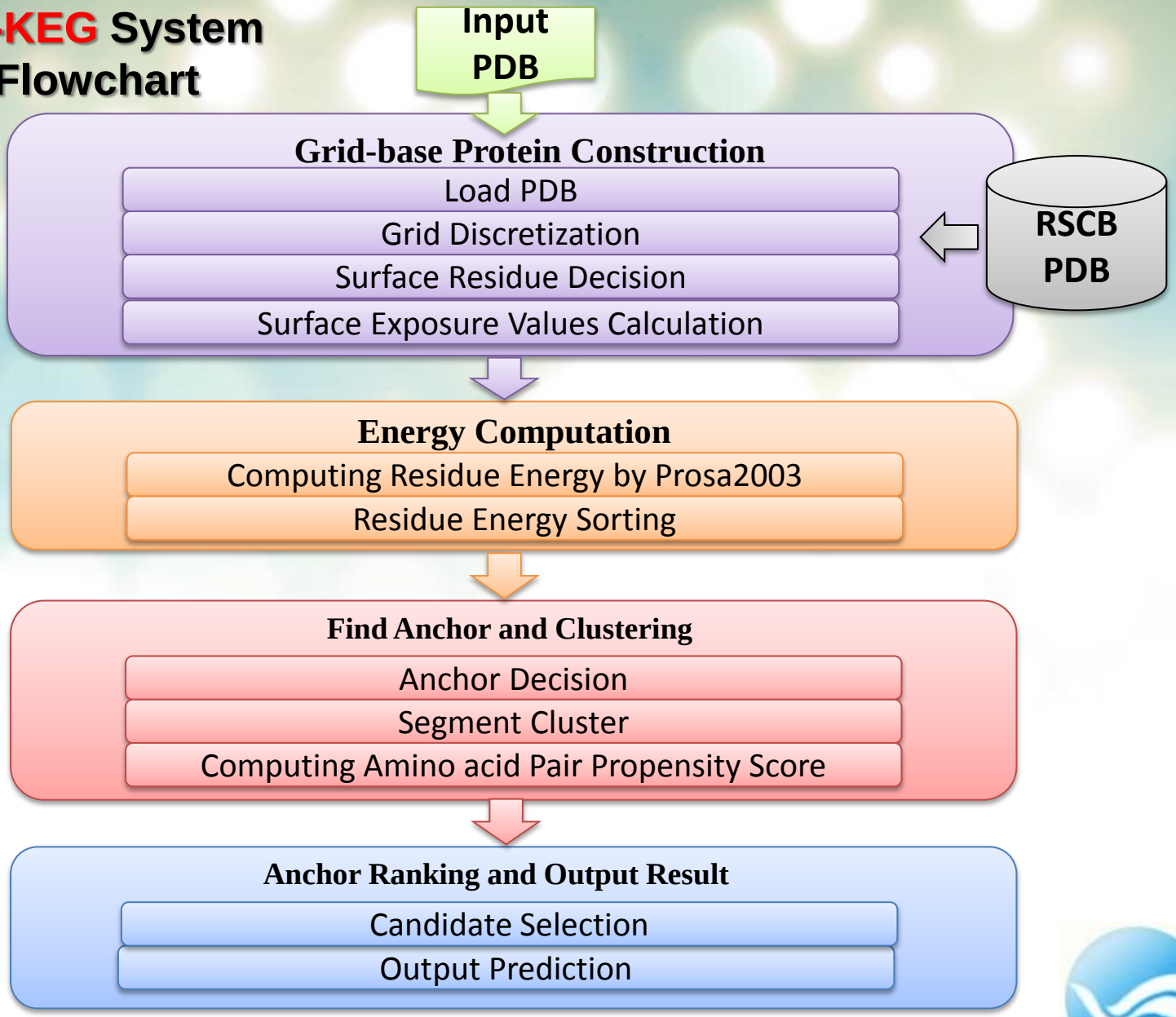
- These prediction tools adopted various combinations of physical-chemical characteristics and trained statistical features from known antigen-antibody complexes to identify CE candidates.

Our GOAL

- To develop a new CE Prediction System using Energy and Residue Contents.

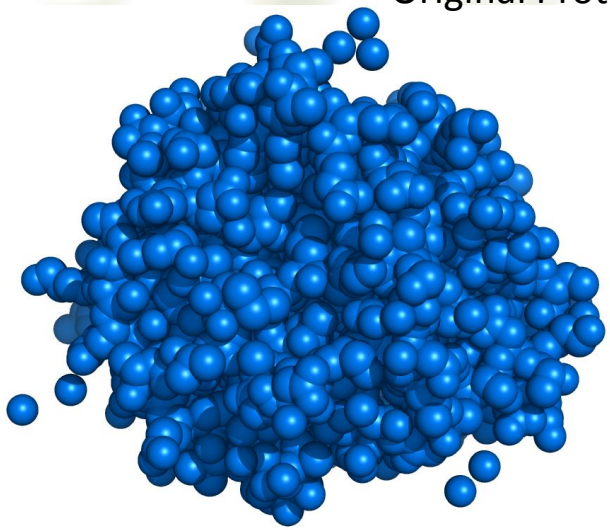


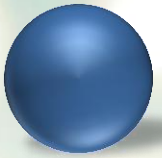
CE-KEG System Flowchart



Grid-base Protein Construction(1/3)

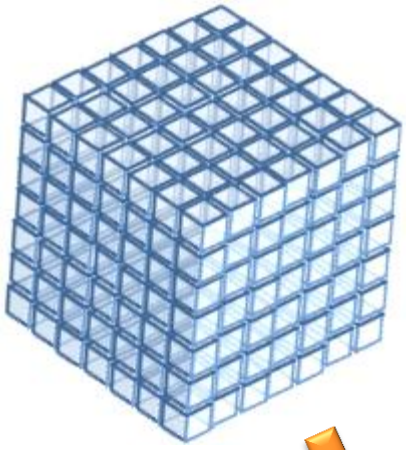
Original Protein

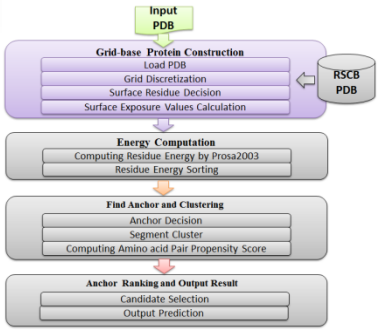


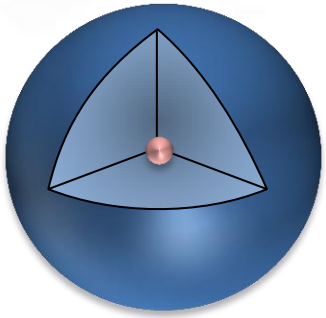


Atom

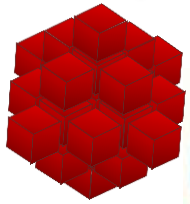
Discretization





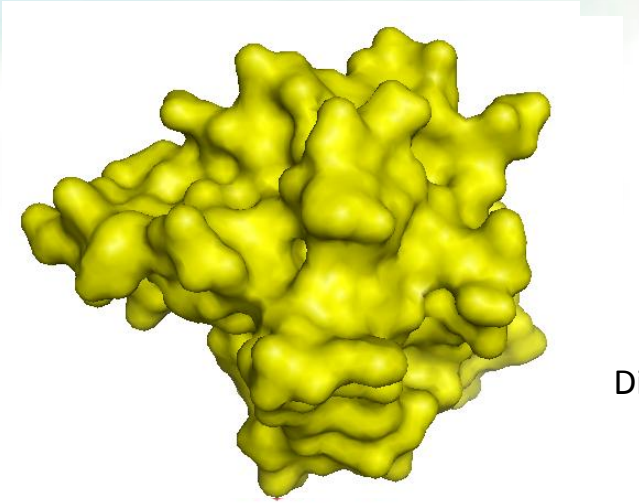


An atom with a corresponding radius



Grid-based atom

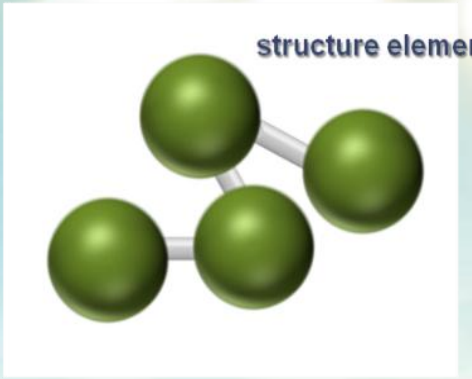
Grid-base Protein Construction(2/3)



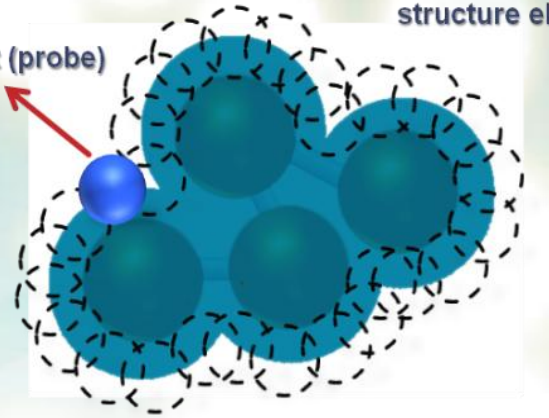
PDB ID: 1ACB

Discretization

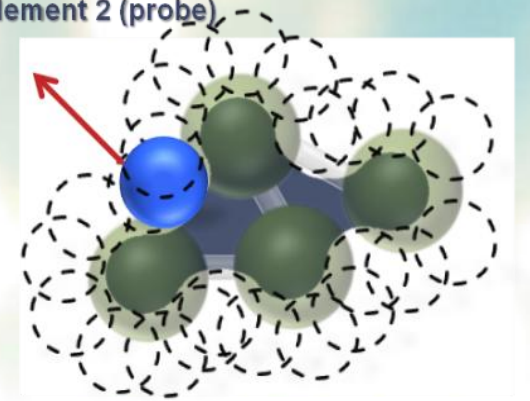
Grid-base Protein Construction(3/3)



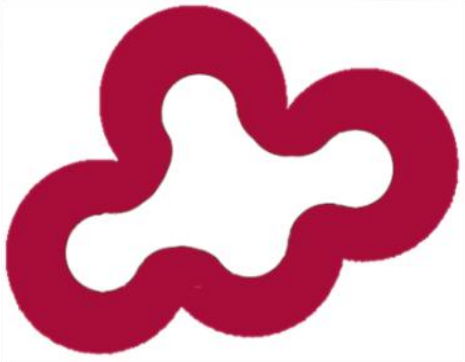
1. original structure



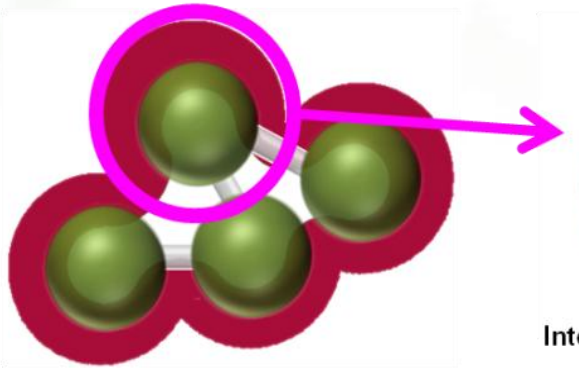
2. Dilated result

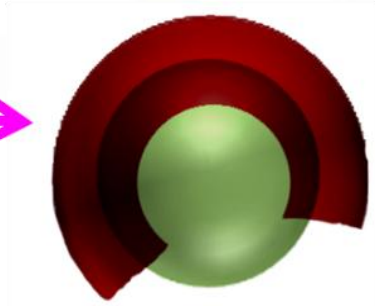


3. Erode the Dilated result



4. Difference result (2 – 3)

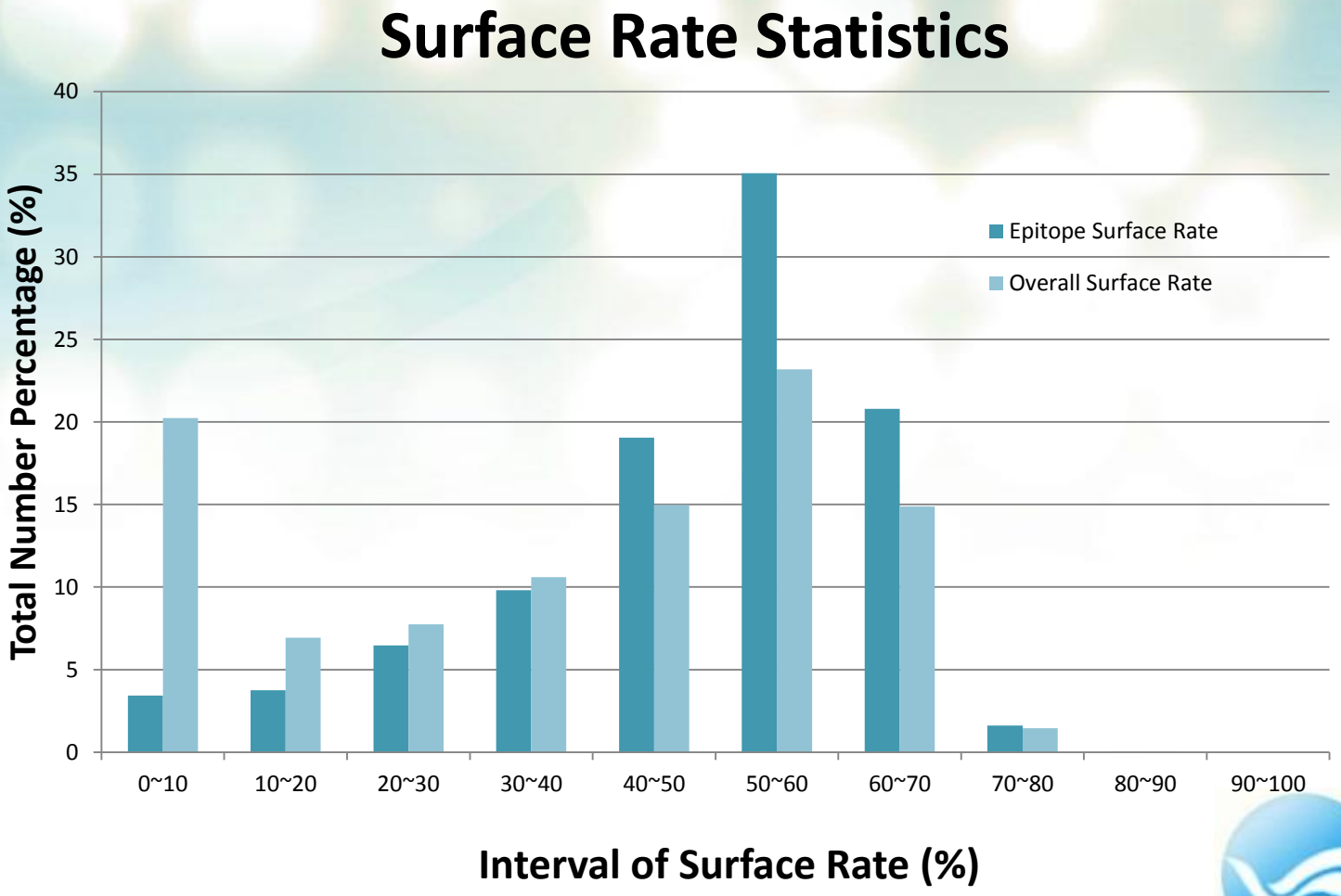


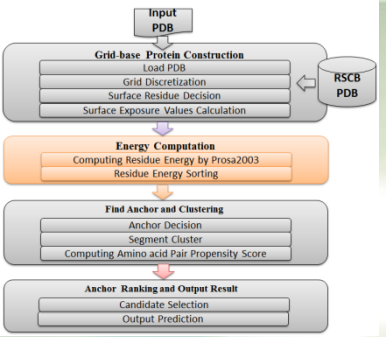


Intersected regions / total regions

5. atom surface rate computation

The distribution of surface rate in true CE Residues and overall residues

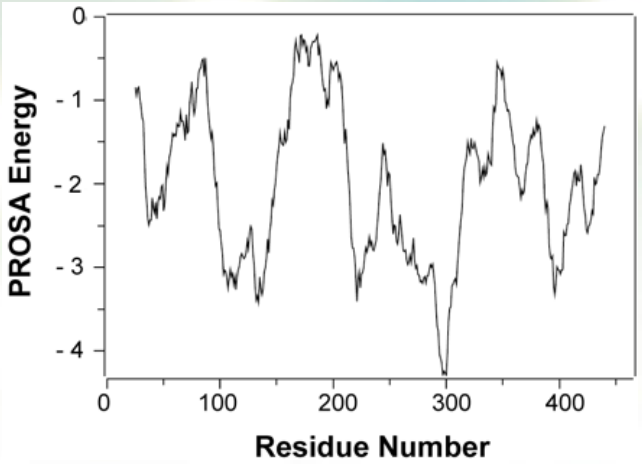




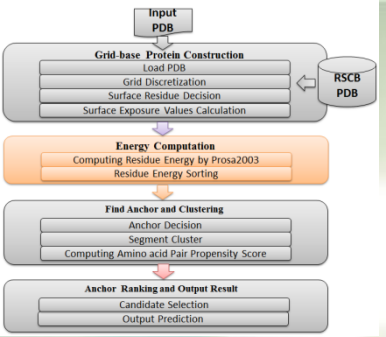
Energy Computation

- Software: Prosa2003
- Function:
  - improving the folding recognition
  - structure prediction and refinement
- Description:
  - The knowledge-based potential was adopted for representing the energy of each surface residue, which was obtained from the distribution of pairwise distances to extract effective potentials between residues.

The result of Prosa2003



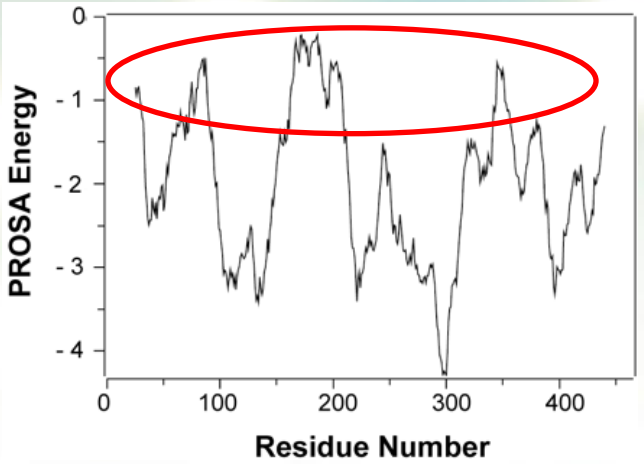
Here we adopted the advantages of calculated energy function from each surface residue to distinguish various statuses of active conditions.

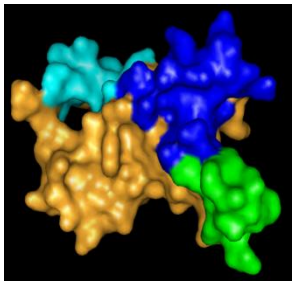


Energy Computation

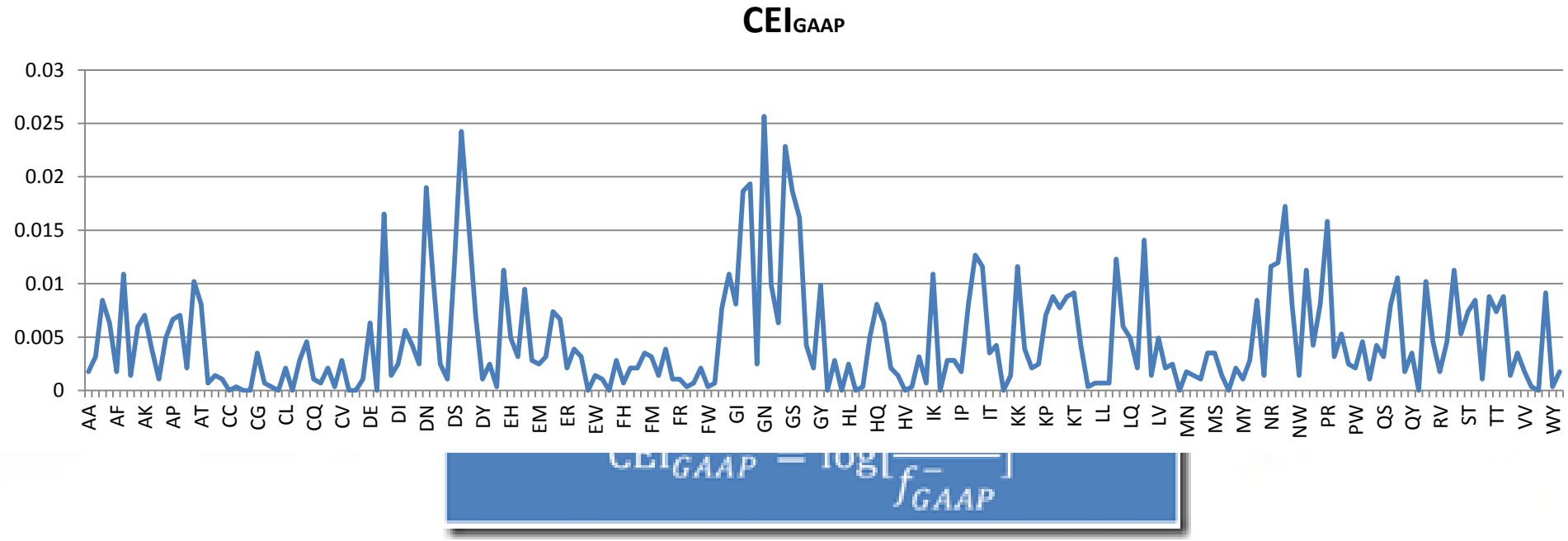
- Step1.Selected the first 20% residues with high energy as our initial CE anchors.
- Step2. Selected initial seeds should possess surface rates.
- Step3. All satisfied seed residues would be mutually examined with a shortest distance of 12 Å to eliminate possible CE candidate groups.
- Step4. the neighboring residues will be included within the radius of 10 Å.

The result of Prosa2003





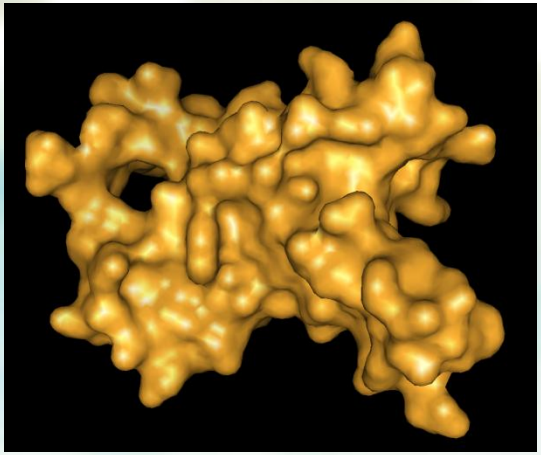
Occurrence Frequency Analysis of Geometrical Amino Acid Pairs



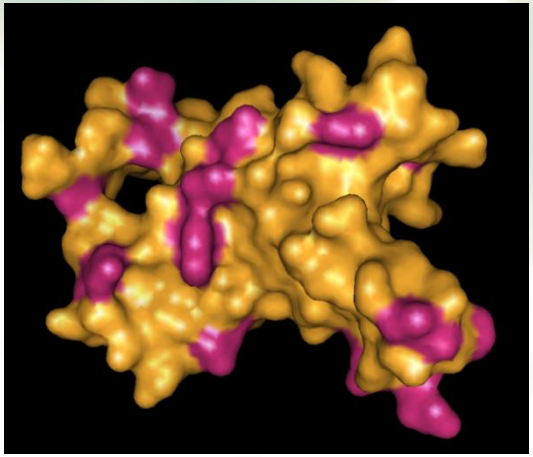
$$CEI_{GAAP} = \log\left[\frac{f_{GAAP}}{\bar{f}_{GAAP}}\right]$$

CE Index (CEI_{GAAP}) : To calculate the frequency of occurrence of a particular pair in the CE dataset divided by the frequency of occurrence of the same pair in the non-CE epitope dataset, and then took logarithm of the ratio to base 10.

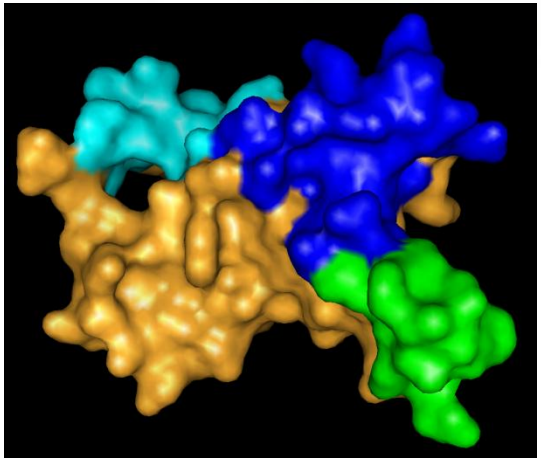
Find Anchor and Clustering



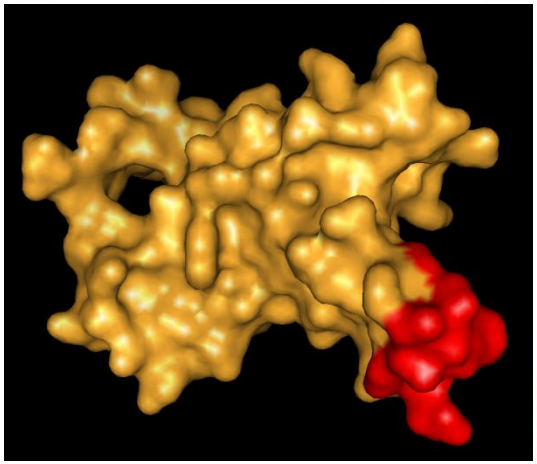
(a)Protein surface detection



(b)energy thresholding



(c)three predicted CE clusters



(d)the true-CE residue of protein 1ORS:C

Experimental Results

Performance measurement

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$PPV = \frac{TP}{TP + FP}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Statistical Results

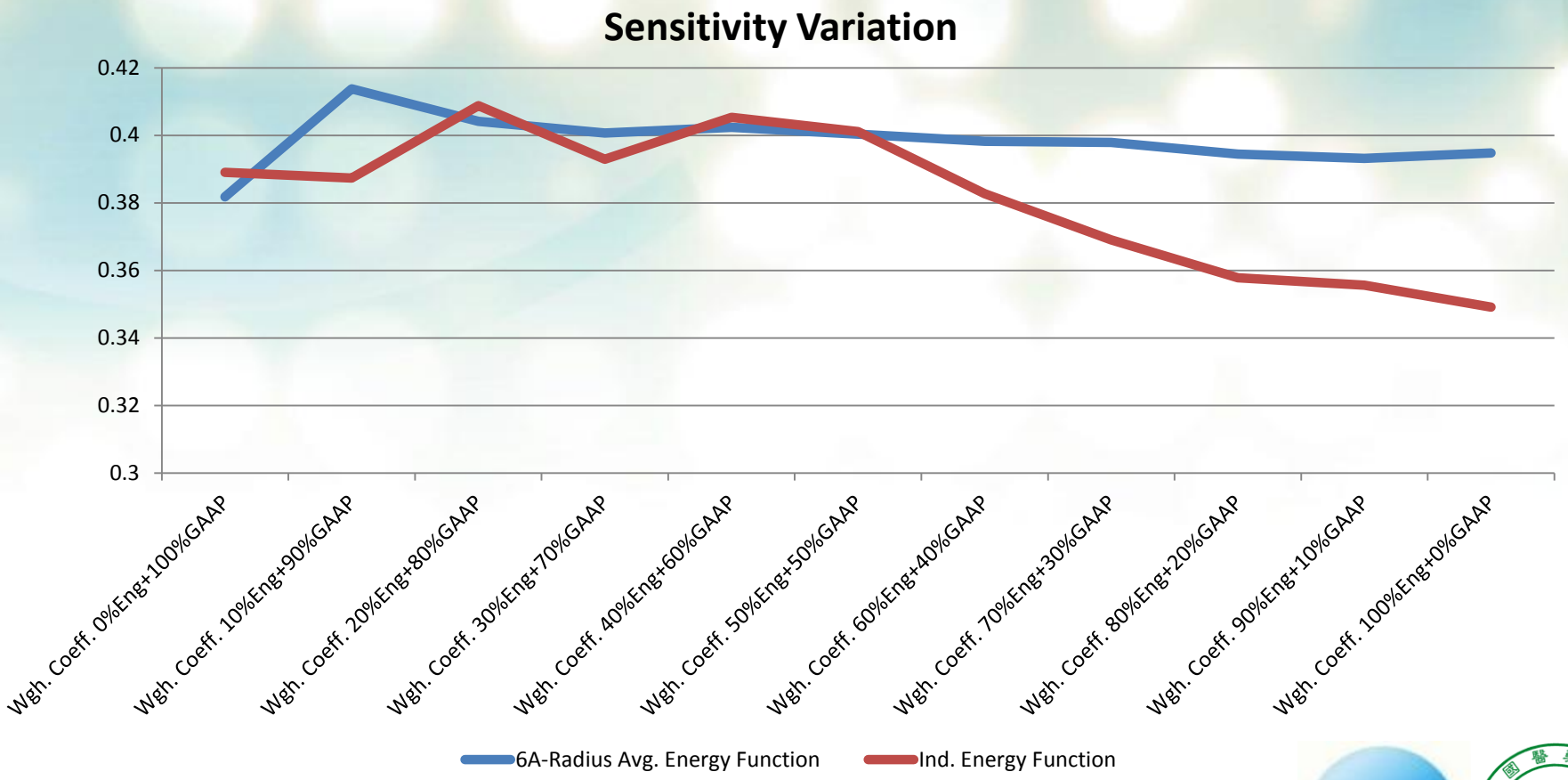
Table 2: Average performance of CE prediction for various weighting coefficient combinations between average energy (Avg. EG) within a 6 Å-radius and pairwise residue occurrence rate (PR). Each antigen was predicted with three CE candidates.

Weighting Combinations	SE	SP	PPV	ACC
0%EG+100%PR	0.38174909	0.88026912	0.28948427	0.82762314
10%EG+90%PR	**0.41375626**	0.88491713	0.318401513	0.83550329
20%EG+80%PR	0.40411907	0.88339643	0.310372011	0.83364651
30%EG+70%PR	0.40071021	0.88472985	0.308931260	0.83462812
40%EG+60%PR	0.40235963	0.88500477	0.308956909	0.83484050
50%EG+50%PR	0.40032410	0.88526988	0.308866524	0.83494350
60%EG+40%PR	0.39826932	0.88709592	0.310329851	0.83674728
70%EG+30%PR	0.39788531	0.88708866	0.310057838	0.83681763
80%EG+20%PR	0.39440495	0.88639840	0.307165993	0.83575056
90%EG+10%PR	0.39315133	0.88647102	0.307463589	0.83588749
100%EG+0%PR	0.39477960	0.88665173	0.307860654	0.83606191

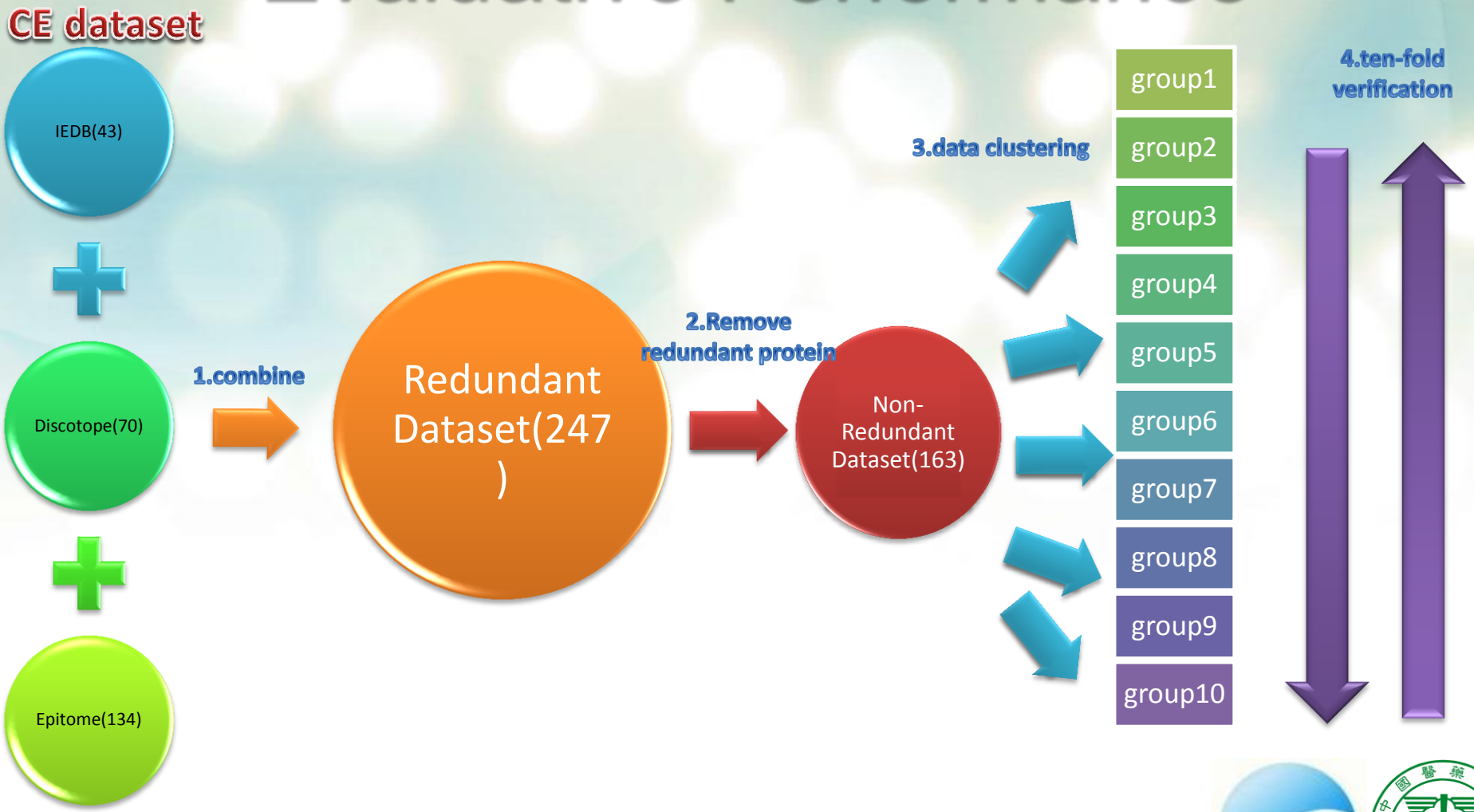
Table 3: Average performance of CE prediction for various weighting coefficient combinations between individual energy (Ind. EG) and pairwise residue occurrence rate (PR). Each antigen was predicted with three CE candidates.

Weighting Combinations	SE	SP	PPV	ACC
0%EG+100%PR	0.38904213	0.88545484	0.297620232	0.83316720
10%EG+90%PR	0.38730979	0.88374611	0.295145236	0.83109301
20%EG+80%PR	**0.40874497**	0.88785200	0.315718499	0.83729001
30%EG+70%PR	0.39293810	0.88612791	0.305437883	0.83393131
40%EG+60%PR	0.40530435	0.88759054	0.313223041	0.83635800
50%EG+50%PR	0.40110938	0.88624436	0.314452191	0.83427900
60%EG+40%PR	0.38267268	0.88614126	0.306830027	0.83289012
70%EG+30%PR	0.36904261	0.88510455	0.297330839	0.83028217
80%EG+20%PR	0.35784993	0.88327931	0.287382221	0.82740505
90%EG+10%PR	0.35565826	0.88242811	0.283611851	0.82639348
100%EG+0%PR	0.349151010	0.88206203	0.281820846	0.82577874

Statistical Results



Evaluative Performance



Evaluative Performance (Ten-fold & Ten Times)

3 dataset CE dataset (Total 248 proteins)

	Ten-fold Results											Average
	Test1	Test2	Test3	Test4	Test5	Test6	Test7	Test8	Test9	Test10		Ten-fold
Sensitivity	0.383160	0.379787	0.37729	0.379258	0.385349	0.388107	0.379098	0.379132	0.386024	0.375141		**0.381235**
Specificity	0.880571	0.881009	0.87914	0.880072	0.87981	0.880345	0.879872	0.887975	0.878763	0.878127		**0.880568**
PPV	0.297399	0.287289	0.288493	0.291191	0.287539	0.296898	0.29054	0.305936	0.289669	0.283264		**0.291822**
Accuracy	0.828234	0.82883	0.82649	0.827769	0.827823	0.828432	0.827397	0.833007	0.826407	0.82499		**0.827938**

Remove redundant data (Total 163 proteins)

	Ten-fold Results											Average
	Test1	Test2	Test3	Test4	Test5	Test6	Test7	Test8	Test9	Test10		Ten-fold
Sensitivity	0.341964	0.348149	0.348203	0.351525	0.341929	0.346977	0.347666	0.347436	0.347639	0.33798		**0.345947**
Specificity	0.886391	0.887023	0.887015	0.884151	0.886151	0.886993	0.887045	0.887073	0.887108	0.891938		**0.887089**
PPV	0.287202	0.292822	0.293277	0.289017	0.286981	0.292107	0.292636	0.292365	0.292624	0.304628		**0.292366**
Accuracy	0.826815	0.827903	0.82769	0.825199	0.826284	0.827708	0.827778	0.827831	0.827852	0.82929		**0.827435**

Conclusions

Conclusions

- In this paper, a novel method combined characteristics of surface rate, energy function, and geometrical amino acid pairs was proposed for predicting CE residues located in discontinuous B cell antigenic determinates.
- To compare the prediction performance with DiscoTope system with respect to the DiscoTope's testing dataset
  - average specificity： **0.891(CE-KEG)** > 0.75 (DiscoTope)
  - average sensitivity： **0.565(CE-KEG)** > 0.473 (DiscoTope)
  - AUC{(spe+sen)/2}： **0.728(CE-KEG)** > 0.621(DiscoTope)
- To compare the prediction performance with PEPITO (BEPro) system
  - with respect to the Epitome's testing dataset
    - AUC{(spe+sen)/2}： **0.694(CE-KEG)** > 0.683(BEPro)
  - with respect to the DiscoTope's testing dataset
    - AUC{(spe+sen)/2}： 0.728(CE-KEG) < **0.753(BEPro)**

Demo

- http://cekeg.ntou.edu.tw

Current Query

- Lorem ipsum dolor sit amet
- Lorem ipsum dolor sit amet
- Lorem ipsum dolor sit amet
- Lorem ipsum dolor sit amet

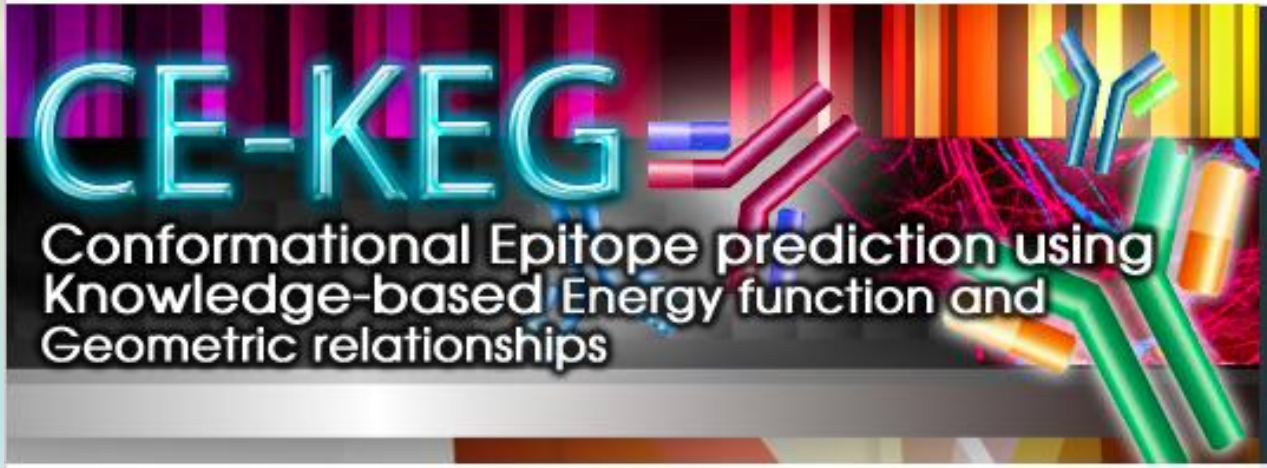
Welcome to our site

CEKEG-Prediction Method

The grid-based and mathematical morphological algorithms were applied for efficient detection and extraction of surface atoms, and initial surface resides of predicted CE candidates were exclusively selected according to the local average energy distribution. The novel CE prediction system was then developed based on the characteristics of surface rates, occurrence frequency of geometrical neighbouring residue combination, and knowledge-based energy functions. The trained and weighted combinatorial features of surface residue contents and potentials were integrated for a simple and effective CE prediction system.

Related Server Link

- **SEPPA server**: With 3D protein structure as input, each residue in the query protein will be given a score according to its neighborhood residues information.Higher score corresponds to higher probability the residue to be involved in an epitope.
- **DiscoTope Server** utilizes calculation of surface accessibility (estimated in terms of contact numbers) and a novel



Current Query

- Lorem ipsum dolor sit amet
- Lorem ipsum dolor sit amet
- Lorem ipsum dolor sit amet
- Lorem ipsum dolor sit amet

Welcome to our site

Enter a PDB ID and its chain ID:

PDB ID: 1acb

or upload a pdbfile:

選擇檔案 尚未選取檔案

Submit Reset

Related Server Link

- **SEPPA server**: With 3D protein structure as input, each residue in the query protein will be given a score according to its neighborhood residues information.Higher score corresponds to higher probability the residue to be involved in an epitope.
- **DiscoTope Server** utilizes calculation of surface accessibility (estimated in terms of contact numbers) and a novel

Center

Current Query

- Lorem ipsum dolor sit amet
- Lorem ipsum dolor sit amet
- Lorem ipsum dolor sit amet
- Lorem ipsum dolor sit amet

Step2:

Select 1acb its chain :

(•) E ( ) I

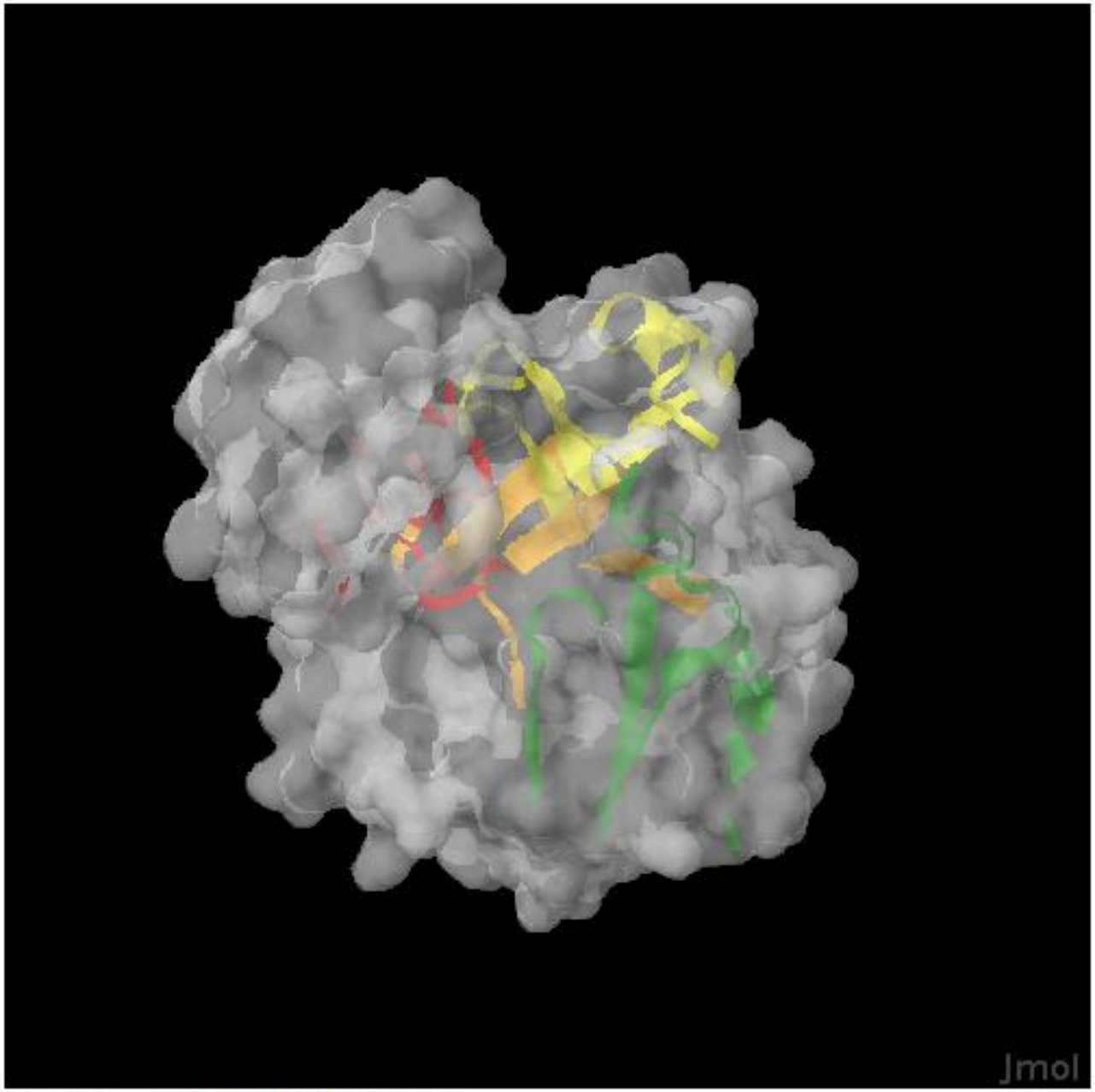
Ok

Related Server Link

- **SEPPA server**: With 3D protein structure as input, each residue in the query protein will be given a score according to its neighborhood residues information.Higher score corresponds to higher probability the residue to be involved in an epitope.

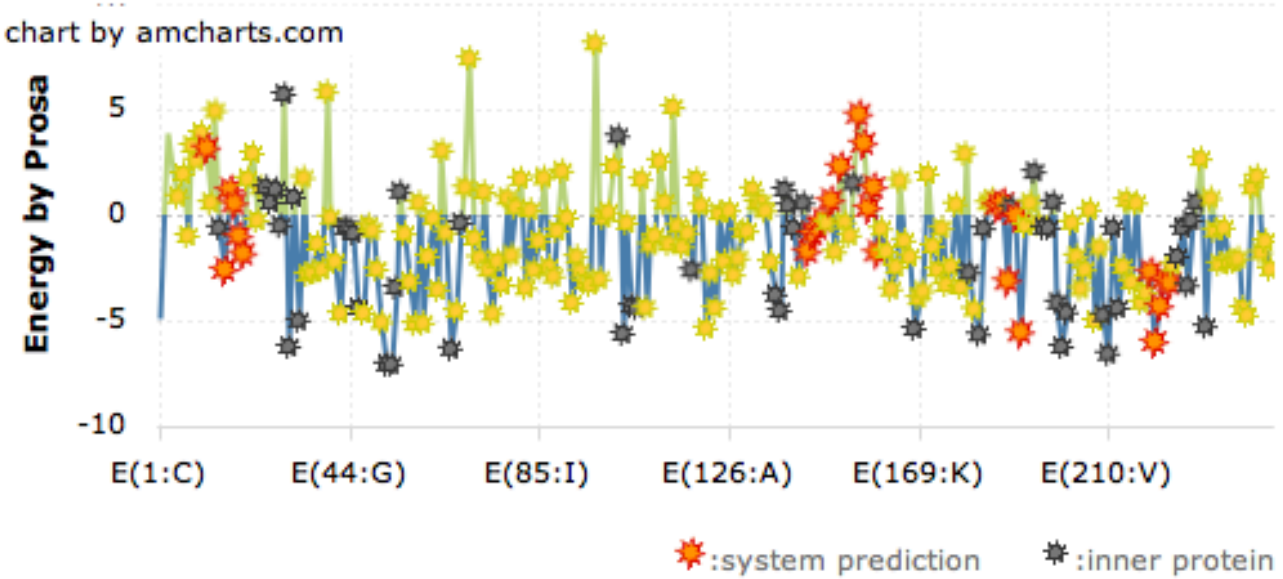
Prediction Result

PDB ID:1ACB | chain ID:E | Residue Number#241

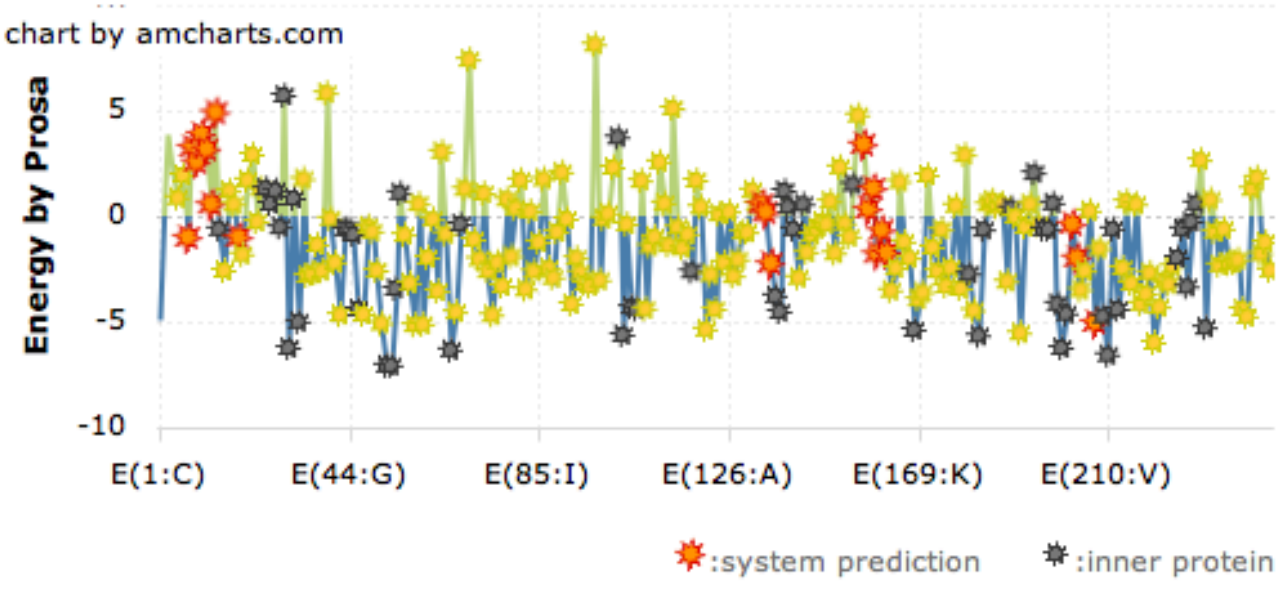


Rank# #all #1 #2 #3 #4

Prediction_1: (Chain: E)	11,17,18,19,20,21,138,140,142,143,144,145,146,150,152,156,157, 158,159,160,186,187,188,189,190,191,194,219,220,221,222,223



Prediction_2: (Chain: E)	7,8,9,10,11,12,13,20,27,133,134,135,136,137,138,157,158,159,16 0,161,162,200,201,202,203,207




thanks for your attention