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

Ying-Tsang Lo¹, Tun-Wen Pai¹,²*, Wei-Kuo Wu¹, Hao-Teng Chang³,⁴*

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

Proportion of native protein

- ~10% CE
- ~90% LE

PDBid: 1DUY

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

Input PDB

Grid-base Protein Construction
- Load PDB
- Grid Discretization
- Surface Residue Decision
- Surface Exposure Values Calculation

RSCB PDB

Energy Computation
- Computing Residue Energy by Prosa2003
- Residue Energy Sorting

Find Anchor and Clustering
- Anchor Decision
- Segment Cluster
- Computing Amino acid Pair Propensity Score

Anchor Ranking and Output Result
- Candidate Selection
- Output Prediction
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)
5. Atom surface rate computation
The distribution of surface rate in true CE Residues and overall residues

Surface Rate Statistics

- Epitope Surface Rate
- Overall Surface Rate

Interval of Surface Rate (%)

Total Number Percentage (%)
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.

*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**.
Occurrence Frequency Analysis of Geometrical Amino Acid Pairs

CE Index (CEI\textsubscript{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.**

<table>
<thead>
<tr>
<th>Weighting Combinations</th>
<th>SE</th>
<th>SP</th>
<th>PPV</th>
<th>ACC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%EG+100%PR</td>
<td>0.38174909</td>
<td>0.88026912</td>
<td>0.28948427</td>
<td>0.82762314</td>
</tr>
<tr>
<td>10%EG+90%PR</td>
<td><strong>0.41375626</strong></td>
<td>0.88491713</td>
<td>0.318401513</td>
<td>0.83550329</td>
</tr>
<tr>
<td>20%EG+80%PR</td>
<td>0.40411907</td>
<td>0.88339643</td>
<td>0.310372011</td>
<td>0.83364651</td>
</tr>
<tr>
<td>30%EG+70%PR</td>
<td>0.40071021</td>
<td>0.88742985</td>
<td>0.308931260</td>
<td>0.83462812</td>
</tr>
<tr>
<td>40%EG+60%PR</td>
<td>0.40235963</td>
<td>0.88500477</td>
<td>0.30956909</td>
<td>0.83484050</td>
</tr>
<tr>
<td>50%EG+50%PR</td>
<td>0.40032410</td>
<td>0.88526988</td>
<td>0.30886524</td>
<td>0.83494350</td>
</tr>
<tr>
<td>60%EG+40%PR</td>
<td>0.39826932</td>
<td>0.88709592</td>
<td>0.310329851</td>
<td>0.83674728</td>
</tr>
<tr>
<td>70%EG+30%PR</td>
<td>0.39788531</td>
<td>0.88708866</td>
<td>0.310057838</td>
<td>0.83681763</td>
</tr>
<tr>
<td>80%EG+20%PR</td>
<td>0.39440495</td>
<td>0.88639840</td>
<td>0.307165993</td>
<td>0.83575056</td>
</tr>
<tr>
<td>90%EG+10%PR</td>
<td>0.39315133</td>
<td>0.88647102</td>
<td>0.307463589</td>
<td>0.83588749</td>
</tr>
<tr>
<td>100%EG+0%PR</td>
<td>0.39477960</td>
<td>0.88665173</td>
<td>0.307860654</td>
<td>0.83606191</td>
</tr>
</tbody>
</table>

**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.**

<table>
<thead>
<tr>
<th>Weighting Combinations</th>
<th>SE</th>
<th>SP</th>
<th>PPV</th>
<th>ACC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%EG+100%PR</td>
<td>0.38904213</td>
<td>0.88545484</td>
<td>0.297620232</td>
<td>0.83316720</td>
</tr>
<tr>
<td>10%EG+90%PR</td>
<td>0.38730979</td>
<td>0.88374611</td>
<td>0.295145236</td>
<td>0.83109301</td>
</tr>
<tr>
<td>20%EG+80%PR</td>
<td><strong>0.40874497</strong></td>
<td>0.88785200</td>
<td>0.315718499</td>
<td>0.83729001</td>
</tr>
<tr>
<td>30%EG+70%PR</td>
<td>0.39293810</td>
<td>0.88612791</td>
<td>0.305437883</td>
<td>0.83393131</td>
</tr>
<tr>
<td>40%EG+60%PR</td>
<td>0.40530435</td>
<td>0.88759054</td>
<td>0.313223041</td>
<td>0.83658000</td>
</tr>
<tr>
<td>50%EG+50%PR</td>
<td>0.40110938</td>
<td>0.88624436</td>
<td>0.314452191</td>
<td>0.83427900</td>
</tr>
<tr>
<td>60%EG+40%PR</td>
<td>0.38267268</td>
<td>0.88614126</td>
<td>0.306830027</td>
<td>0.83289012</td>
</tr>
<tr>
<td>70%EG+30%PR</td>
<td>0.36904261</td>
<td>0.88510455</td>
<td>0.297330839</td>
<td>0.83028217</td>
</tr>
<tr>
<td>80%EG+20%PR</td>
<td>0.35784993</td>
<td>0.88327931</td>
<td>0.287382221</td>
<td>0.82740505</td>
</tr>
<tr>
<td>90%EG+10%PR</td>
<td>0.35565826</td>
<td>0.88242811</td>
<td>0.283611851</td>
<td>0.82639348</td>
</tr>
<tr>
<td>100%EG+0%PR</td>
<td>0.349151010</td>
<td>0.88206203</td>
<td>0.281820846</td>
<td>0.82577874</td>
</tr>
</tbody>
</table>
Evaluative Performance

1. Combine

IEDB(43)
Discotope(70)
Epitome(134)

2. Remove redundant protein

Redundant Dataset(247)

3. Data clustering

Non-Redundant Dataset(163)

group1
group2
group3
group4
group5
group6
group7
group8
group9
group10

4. Ten-fold verification

CE dataset

NTOU CSE GENOMICS & BIOINFORMATICS LAB
# Evaluative Performance

(Ten-fold & Ten Times)

<table>
<thead>
<tr>
<th>3 dataset</th>
<th>CE dataset (Total 248 proteins)</th>
</tr>
</thead>
</table>

### Ten-fold Results

<table>
<thead>
<tr>
<th></th>
<th>Test1</th>
<th>Test2</th>
<th>Test3</th>
<th>Test4</th>
<th>Test5</th>
<th>Test6</th>
<th>Test7</th>
<th>Test8</th>
<th>Test9</th>
<th>Test10</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sensitivity</strong></td>
<td>0.383160</td>
<td>0.379787</td>
<td>0.37729</td>
<td>0.379258</td>
<td>0.385349</td>
<td>0.388107</td>
<td>0.379098</td>
<td>0.379132</td>
<td>0.386024</td>
<td>0.375141</td>
<td><strong>0.381235</strong></td>
</tr>
<tr>
<td><strong>Specificity</strong></td>
<td>0.880571</td>
<td>0.881009</td>
<td>0.87914</td>
<td>0.880072</td>
<td>0.879872</td>
<td>0.880345</td>
<td>0.87975</td>
<td>0.878763</td>
<td>0.878127</td>
<td></td>
<td><strong>0.880568</strong></td>
</tr>
<tr>
<td><strong>PPV</strong></td>
<td>0.297399</td>
<td>0.287289</td>
<td>0.288493</td>
<td>0.291191</td>
<td>0.287539</td>
<td>0.296898</td>
<td>0.29054</td>
<td>0.305936</td>
<td>0.289669</td>
<td>0.283264</td>
<td><strong>0.291822</strong></td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>0.828234</td>
<td>0.82883</td>
<td>0.82649</td>
<td>0.827769</td>
<td>0.827823</td>
<td>0.828432</td>
<td>0.827397</td>
<td>0.833007</td>
<td>0.826407</td>
<td>0.82499</td>
<td><strong>0.827938</strong></td>
</tr>
</tbody>
</table>

---

### Ten-fold Results

<table>
<thead>
<tr>
<th></th>
<th>Test1</th>
<th>Test2</th>
<th>Test3</th>
<th>Test4</th>
<th>Test5</th>
<th>Test6</th>
<th>Test7</th>
<th>Test8</th>
<th>Test9</th>
<th>Test10</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sensitivity</strong></td>
<td>0.331964</td>
<td>0.348149</td>
<td>0.348203</td>
<td>0.351525</td>
<td>0.341929</td>
<td>0.346977</td>
<td>0.347666</td>
<td>0.347436</td>
<td>0.347639</td>
<td>0.33798</td>
<td><strong>0.345947</strong></td>
</tr>
<tr>
<td><strong>Specificity</strong></td>
<td>0.886391</td>
<td>0.887023</td>
<td>0.887015</td>
<td>0.884151</td>
<td>0.886151</td>
<td>0.886993</td>
<td>0.887045</td>
<td>0.887073</td>
<td>0.887108</td>
<td>0.891938</td>
<td><strong>0.887089</strong></td>
</tr>
<tr>
<td><strong>PPV</strong></td>
<td>0.287202</td>
<td>0.292822</td>
<td>0.293277</td>
<td>0.289017</td>
<td>0.286981</td>
<td>0.292107</td>
<td>0.292636</td>
<td>0.292365</td>
<td>0.292624</td>
<td>0.304628</td>
<td><strong>0.292366</strong></td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>0.826815</td>
<td>0.827903</td>
<td>0.82769</td>
<td>0.825199</td>
<td>0.826284</td>
<td>0.827708</td>
<td>0.827778</td>
<td>0.827831</td>
<td>0.827852</td>
<td>0.82929</td>
<td><strong>0.827435</strong></td>
</tr>
</tbody>
</table>

---

Remove redundant data (Total 163 proteins)
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: \(0.891(\text{CE-KEG}) > 0.75\) (DiscoTope)
  – average sensitivity: \(0.565(\text{CE-KEG}) > 0.473\) (DiscoTope)
  – AUC\(((\text{spe}+\text{sen})/2)\): \(0.728(\text{CE-KEG}) > 0.621\) (DiscoTope)

• To compare the prediction performance with PEPITO (BEPro) system - with respect to the Epitome’s testing dataset
  • AUC\(((\text{spe}+\text{sen})/2)\): \(0.694(\text{CE-KEG}) > 0.683\) (BEPro)

- with respect to the DiscoTope’s testing dataset
  • AUC\(((\text{spe}+\text{sen})/2)\): \(0.728(\text{CE-KEG}) < 0.753\) (BEPro)
Demo

- http://cekeg.ntou.edu.tw
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 residues 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
Welcome to our site

Enter a PDB ID and its chain ID:

PDB ID: 1acb

or upload a pdb file:

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
CE-KEG
Conformational Epitope prediction using Knowledge-based Energy function and Geometric relationships

Step 2:

Select 1acb its chain:

[E][I] Ok

Current Query

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

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
**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

**chart by amcharts.com**

- **Energy by Prosa**

- :system prediction
- *:inner protein

**Prediction 2:**
(Chain: E)
7,8,9,10,11,12,13,20,27,133,134,135,136,137,138,157,158,159,160,161,162,200,201,202,203,207

**chart by amcharts.com**

- **Energy by Prosa**

- :system prediction
- *:inner protein
thanks for your attention