A Multimodal Deep Architecture for Large-Scale Protein Ubiquitylation Site Prediction

Fei He¹, Lingling Bao¹, Rui Wang¹, Jiagen Li¹, Dong Xu², Xiaowei Zhao¹

¹School of Information Science and Technology
Institute of Computational Biology
Northeast Normal University Changchun, China

²Department of Electrical Engineering and Computer Science, Informatics Institute, and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA
The conjugation of ubiquitin to a substrate protein on a particular lysine involves three types of enzymes (activating, ligases and conjugatin enzymes) related to various cellular functions.
Introduction

Experimental techniques
- Laborious
- Expensive
- Time consuming
- Mass spectrometry
- CHIP-CHIP analysis

Computational approaches
- Convenient
- Efficient
- Guidance
- ESA-Ubisite
- UbiProber
- iUbiq-Lys
- Ubisite
### Introduction

<table>
<thead>
<tr>
<th>Tool</th>
<th>Metrics</th>
<th>Testing data scale (sites)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Sensitivity</td>
</tr>
<tr>
<td><strong>ESA-UbiSite</strong>&lt;sup&gt;1&lt;/sup&gt;</td>
<td>92%</td>
<td>66%</td>
</tr>
<tr>
<td></td>
<td>61.26%</td>
<td>46.14%</td>
</tr>
<tr>
<td><strong>UbiProber</strong>&lt;sup&gt;2&lt;/sup&gt;</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>55.06%</td>
<td>62.40%</td>
</tr>
<tr>
<td><strong>iUbiq-Lys</strong>&lt;sup&gt;3&lt;/sup&gt;</td>
<td>90.06%</td>
<td>80.99%</td>
</tr>
<tr>
<td></td>
<td>84.63%</td>
<td>3.35%</td>
</tr>
<tr>
<td><strong>UbiSite</strong>&lt;sup&gt;4&lt;/sup&gt;</td>
<td>73.69%</td>
<td>85.10%</td>
</tr>
<tr>
<td></td>
<td>73.63%</td>
<td>29.62%</td>
</tr>
</tbody>
</table>

Introduction

1. Weakness of handcrafted features
   The conventional feature engineering leads to produce biased and incomplete features

2. Heterogeneity among different features
   (inspired by iris features) Most existed tools combine multiple modal features

3. Extreme imbalanced data problem
   Only a small size of lysine can be attached to ubiquitin

Challenges on large-scale protein ubiquitylation site prediction
Introduction

Detect underlying informative patterns in protein sequence

Deep network

Multi-modal fusion

Involve three types of informative modalities and fuse their generative deep representations

Bootstrap strategy

Control the model unbiased and stable on imbalanced data
Remove redundancy
Three different categories of protein modalities

- **One-Hot vector**
- physico-chemical properties
- position-specific scoring matrix (PSSM)

**Information Sources**

- Sequence information
- PCP for amino acid
- Evolutionary profiles
Each amino acid is encoded into a 21-dimensional binary vector.

The index corresponding to the amino acid is 1, and other positions are 0.

A dash will be filled in absent positions and be encoded to 0.05 (1/21).
● There is a strong connection between physico-chemical properties of amino acids and ubiquitylation sites\textsuperscript{1,2}

● Select top 13 physico-chemical properties in light of the literature\textsuperscript{3}

\begin{table}
\centering
\begin{tabular}{|c|c|p{8cm}|}
\hline
No & PCP & Description \\
\hline
1 & EISD860102 & Atom-based hydrophobic moment \\
2 & ZIMJ680104 & Isoelectric point \\
3 & HUTJ700103 & Entropy of formation \\
4 & KARP850103 & Flexibility parameter for two rigid neighbors \\
5 & JANJ780101 & Average accessible surface area \\
6 & FAUJ880111 & Positive charge \\
7 & GUYH850104 & Apparent partition energies calculated from Janin index \\
8 & JANJ780103 & Percentage of exposed residues \\
9 & JANJ790102 & Transfer free energy \\
10 & PONP800102 & Average gain in surrounding hydrophobicity \\
11 & CORJ870101 & NNEIG index \\
12 & VINM940101 & Normalized flexibility parameters, average \\
13 & OOBM770101 & Average non-bonded energy per atom \\
\hline
\end{tabular}
\end{table}


Demonstrate the evolutionary profiles of the protein sequences

Search database: Swiss-Prot

Parameter: num_iterations 3, e-value 0.001

Logistic normalization: \( \frac{1}{1 + e^{-x}} \)
Method

One hot vector

3-layer 1D CNN to detect **local structural feature maps**

3-layer DNN to combine all feature maps

**Physico-chemical properties**

3-layer DNN to **interconnect all properties** for their joint effect

**PSSM**

3-layer 1D CNN to detect **evolutionary profile among amino acids**

Another 3-layer 1D CNN to detect **evolutionary information cross positions from trans-positional PSSM**

2-layer DNN to merge the feature maps
Run on: Theano 0.9, Keras 1.1.0, GTX1080Ti

Bootstrapping

\[ N = \frac{\text{Neg}}{\text{Pos}} \]

Pre-train

One-hot vector sub-net

Physico-chemical Properties sub-net

PSSM sub-net

Merged Model

Fine-tune
3

Material

12,100 protein sequences with 54,586 Ubiquitylation sites and 320,083 non-ubiquitylation sites

1345 proteins with 6293 Ubiquitylation sites and 46,080 non-ubiquitylation sites

30% of training set extracted as validation samples

Protein Lysine Modification Database

The available largest scale protein ubiquitylation dataset. Extended from CPLA 1.0 and CPLM 2.0 Never mentioned in any other protein ubiquitylation site prediction research.
## Method

<table>
<thead>
<tr>
<th>Subnet</th>
<th>Layer</th>
<th>Activation function</th>
<th>Size</th>
<th>Filters</th>
<th>Dropout</th>
</tr>
</thead>
<tbody>
<tr>
<td>One hot vector</td>
<td>1D Convolution</td>
<td>softsign</td>
<td>2</td>
<td>200</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>softsign</td>
<td>3</td>
<td>150</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>softsign</td>
<td>5</td>
<td>150</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>softsign</td>
<td>7</td>
<td>100</td>
<td>0.4</td>
</tr>
<tr>
<td>Dense</td>
<td>relu</td>
<td>256</td>
<td>--</td>
<td>--</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>relu</td>
<td>128</td>
<td>--</td>
<td>--</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>relu</td>
<td>128</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subnet</th>
<th>Layer</th>
<th>Activation function</th>
<th>Size</th>
<th>Filters</th>
<th>Dropout</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physico-chemical properties</td>
<td>Dense</td>
<td>softplus</td>
<td>1024</td>
<td>--</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>softplus</td>
<td>512</td>
<td>--</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>softplus</td>
<td>256</td>
<td>--</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>relu</td>
<td>128</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subnet</th>
<th>Layer</th>
<th>Activation function</th>
<th>Size</th>
<th>Filters</th>
<th>Dropout</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSSM profile</td>
<td>1D Convolution</td>
<td>relu</td>
<td>1</td>
<td>200</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>relu</td>
<td>8</td>
<td>150</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>relu</td>
<td>9</td>
<td>200</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>relu</td>
<td>1</td>
<td>200</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>relu</td>
<td>3</td>
<td>150</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>relu</td>
<td>7</td>
<td>200</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>relu</td>
<td>128</td>
<td>--</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>relu</td>
<td>128</td>
<td>--</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
The accuracy of validation samples using different window sizes on three modalities

The most suitable window size is 49
merged model achieved the better AUC and mean precision than uni-modality

→ make all input modalities at full capacity

one hot vector performed the best among the three input modalities

→ detect underlying expressions from raw protein sequence fragments
Results

t-SNE visualization

Multi-layer CNN

Multi-layer DNN

Physico-chemical properties

One hot vector

PSSM

PSSM-transposition

Merged presentation

Softmax

Output

Positive samples

Negative samples

t-SNE visualization
## Results

Comparative results with other protein Ubiquitylation site prediction tools on PLMD

<table>
<thead>
<tr>
<th>Tool</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESA-Ubisite</td>
<td>61.26%</td>
<td>46.14%</td>
<td>63.34%</td>
<td>0.064</td>
</tr>
<tr>
<td>UbiProber</td>
<td>55.06%</td>
<td>62.40%</td>
<td>54.05%</td>
<td>0.107</td>
</tr>
<tr>
<td>iUbiq-Lys</td>
<td>84.63%</td>
<td>3.35%</td>
<td>96.88%</td>
<td>0.005</td>
</tr>
<tr>
<td>Ubisite</td>
<td>73.63%</td>
<td>29.62%</td>
<td>79.64%</td>
<td>0.073</td>
</tr>
<tr>
<td>Our deep architecture</td>
<td>66.43%</td>
<td>66.67%</td>
<td>66.40%</td>
<td>0.221</td>
</tr>
</tbody>
</table>
**Results**

The ROC and precision-recall curves comparing proposed deep architecture and two other protein ubiquitylation site prediction tools.

- **Receiver Operating Characteristic**
- **Precision-Recall curve**

Only under a certain minor recall, Ubisite achieved higher precision.

- Our model performed at a higher ROC.
- Our model obtained better AUC and mean precision.
- Our deep architecture has evident overall advantages.
Encode each sample into three informative modalities including one hot vector, physico-chemical properties and PSSM.

Establish a multimodal deep architecture fusing these encoding modalities was for robust classification.

Experimental results have proved our effectiveness on the available largest scale data PLMD.

The success of our method is mainly due to the data-driven feature detection in deep learning, the multimodal fusion of deep representations, and the bootstrapping algorithm.
Thanks for listening

Source codes: https://github.com/jiagenlee/deepUbiquitylation
hef740@nenu.edu.cn