Poster | 33.Poster

Wed. Jul 1, 2026 12:10 PM - 1:00 PM JST | Wed. Jul 1, 2026 3:10 AM - 4:00 AM UTC | Poster 1(4FFoyer+401+402)
[P005] Poster

12:10 PM - 1:00 PM JST | 3:10 AM - 4:00 AM UTC

[P03-387] Prediction of sequence-dependent scores reflecting cleavage efficiency for proteases using deep learning

○Bo Zhu[1], Yinghui Yang[1], Keisuke Mizutani[1], Ken Kobayashi[1], Takanobu Yasuda[1], Tetsuya Kitaguchi[1] (1. Institute of Science Tokyo (Japan))

Keywords：protease、protein language model、streptopain、ADAM17

Proteases such as Streptopain (SpeB) and ADAM17 contribute to disease progression through broad and sequence-dependent substrate cleavage. SpeB promotes bacterial invasion, and ADAM17 regulates inflammatory and signaling pathways via ectodomain shedding. Accurate prediction of protease cleavage efficiency is therefore important for understanding related diseases. However, conventional approaches using techniques such as LC-MS/MS are limited in coverage and throughput. Here, we retrained a BERT-based protein language model to predict sequence-dependent cleavage, leveraging its capacity to capture latent sequence features and contextual dependencies [1–2].

Next-generation sequencing (NGS) data from a protease motif mining study [3] were used to achieve amino acid sequences with occurrence counts. After optimizing input sequence length and occurrence distribution for training (50–300 k sequences), the predicted counts by the models SpeB-BERT and ADAM17-BERT showed a Pearson's correlation coefficient of 0.82 and 0.51, respectively, with the NGS counts on the evaluation datasets. The model-based and NGS-based cleavage scores are calculated as the sums of sub-pentamer counts from the model output and the NGS results, respectively, because higher scores reflect greater cleavage potential.

We next applied the models to predict SpeB and ADAM17 cleavage scores for octapeptides extracted from the whole human proteome (hProt-8mers). To validate the performance of the model-based scores in hProt-8mers, we confirmed that of previously characterized substrates. Gasdermin A, an immune defense protein, was ranked within the top 3% of all hProt-8mers by SpeB, and TNF-α, the primary ADAM17 substrate, was ranked within the top 1%. These results support the feasibility of our model for discovering unknown cleavage sites of SpeB and ADAM17 in known proteins. In future studies, hProt-8mers with top 5% scores will be subjected to additional filtering, including subcellular localization and accessibility, followed by experimental validation of cleavage.

Finally, to evaluate whether the predicted cleavage scores reflect actual cleavage efficiency, SpeB was used as a representative case, owing to the availability of cleavage efficiencies for twenty-one substrates [4]. The model-based scores showed a strong correlation with reported cleavage efficiency across twenty-one SpeB substrates (Spearman's $\rho = 0.81$), while NGS-based scores were unavailable for ten substrates due to insufficient sequence coverage. We believe that our models will contribute to elucidating protease-related pathogenesis, designing protease inhibitors, and developing sensing peptide probes for biomedical applications.

References: [1] IEEE Trans Pattern Anal Mach Intell, 2022, 44, 7112–7127. [2] JACS Au, 2025, 5, 955–964. [3] bioRxiv, 2024. 11. 06. 622033. [4] J Biol Chem, 2001, 276, 44551–44556.