A comparative study on peptide encodings for biomedical classification

Sebastian Spänig, Dominik Heider

Department of Bioinformatics, Faculty of Mathematics and Computer Science, University of Marburg, Germany

The key mechanism of many diseases can be encountered on protein level, including cancer, HIV, and neurodegenerative diseases. Moreover, cases involving multi-resistant pathogens have increased to a threatening level. Thus, in the last decades computational biologists developed encodings and machine learning models for automated and accurate classification of peptides and proteins. By engaging research with this respect, even the pharmaceutical industry acknowledges the potential yield from gained insights to medical treatments (Mahlapuu et al., 2016).

Since these models require a fixed-length, numerical input, an essential part of the classification pipeline involves the engineering of representative encodings of the peptide sequence. To this end, several sequence-based encodings have been introduced so far as part of antimicrobial peptides classification studies, including sparse, compositional, and physicochemical encodings (Veltri et al., 2017). In contrast, others explored informative descriptors of the secondary structure, mainly to tackle structure related issues, e.g., the prediction of HIV-1 co-receptor tropism (Löchel et al., 2018).

However, an appropriate literature search results in a variety of promising encodings, leading to the question, which encoding is suitable for a particular classification task and which of the encodings performs better on a particular data set, respectively. Consequently, we compared the performance of state-of-the-art amino acid encodings on independant, biomedical data sets. Due to their high impact, the data encompasses, e.g., antimicrobial peptides as well as HIV-1 co-receptor tropism, but also biological data, such as protein-protein interaction and cell-penetrating peptides, to examine the generalization capabilities of the proposed encodings.

By collecting and testing available encodings, we incur the time-consuming literature search and, in particular, the feature selection. This will aid computational biologist to focus on the actual results instead of occupying themself with choosing an appropriate encoding beforehand. Finally, the overall machine learning part is simplified. In conclusion, our findings will significantly promote classification accuracy specifically for active peptides as well as proteins in general.

References

Löchel, H. F., Riemenschneider, M., Frishman, D., and Heider, D. (2018). SCOTCH: subtype A coreceptor tropism classification in HIV-1. Bioinformatics 34, 2575–2580.

Mahlapuu, M., Håkansson, J., Ringstad, L., and Björn, C. (2016). Antimicrobial Peptides: An Emerging Category of Therapeutic Agents. Front. Cell. Infect. Microbiol. 6, 194.

Veltri, D., Kamath, U., and Shehu, A. (2017). Improving Recognition of Antimicrobial Peptides and Target Selectivity through Machine Learning and Genetic Programming. IEEE/ACM Trans. Comput. Biol. Bioinform. 14, 300–313.