Supplementary Information

Signal Peptide Prediction Using Machine Learning Approaches of Different Complexities

Elias Kahl*, Maximilian Mair*, Julius Schlensok* and Selin Türkoglu*

* all authors contributed equally.
TUM (Technical University of Munich) Department of Informatics, Bioinformatics & Computational Biology - i12, Boltzmannstr. 3, 85748 Garching bei München, Germany

AAindex ID(s)	Feature description	Original publication
KLEP840101	Net charge	Klein et al. (1984). Prediction of protein function from sequence properties. Discriminant analysis of a data base. Biochimica et biophysica acta, https://doi.org/10.1016/0167-4838(84)90312-1
KYTJ820101	Hydropathy index	Kyte & Doolittle (1982). A simple method for displaying the hydropathic character of a protein. Journal of molecular biology https://doi.org/10.1016/0022-2836(82)90515-0
RADA880108	Mean polarity	Radzicka et al. (1988). Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. Biochemistry https://doi.org/10.1021/bi00405a042
FAUJ880103	Normalized van der Waals volume	Fauchère et al. (1988). Amino acid side chain parameters for correlation studies in biology and pharmacology. International journal of peptide and protein research https://doi.org/10.1111/j.1399-3011.1988.tb01261.x
MITS020101	Amphiphilicity index	Mitaku et al. (2002). Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. Bioinformatics (Oxford, England)
CHOP780201, CHOP780202, CHOP780203	Normalized frequency of alpha-helix, Normalized frequency of beta-sheet, Normalized frequency of beta-turn	Chou et al. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. Advances in enzymology and related areas of molecular biology https://doi.org/10.1002/9780470122921.ch2
CHAM810101	Steric parameter	Charton (1981). Protein folding and the genetic code: an alternative quantitative model. Journal of theoretical biology https://doi.org/10.1016/0022-5193(81)90377-5
CHAM830107	A parameter of charge transfer capability	Charton M. & Charton B. I. (1983). The dependence of the Chou-Fasman parameters on amino acid side chain structure. Journal of theoretical biology https://doi.org/10.1016/0022-5193(83)90265-5
JANJ780101	Average accessible surface area	Janin & Wodak (1978). Conformation of amino acid side-chains in proteins. Journal of molecular biology https://doi.org/10.1016/0022-2836(78)90408-4
MEIH800103	Average side chain orientation angle	Meirovitch et al. (1980). Empirical Studies of Hydrophobicity. 1. Effect of Protein Size on the Hydrophobic Behavior of Amino Acids. Macromolecules https://doi.org/10.1021/ma60078a013
VELV850101	Electron-ion interaction potential	Veljković et al. (1985). Is it possible to analyze DNA and protein sequences by the methods of digital signal processing?. IEEE transactions on bio-medical engineering https://doi.org/10.1109/TBME.1985.325549
WERD780101	Propensity to be buried inside	Wertz & Scheraga (1978). Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule. Macromolecules https://doi.org/10.1021/ma60061a002
ZIMJ680105	RF rank	Wertz & Scheraga (1978). Zimmerman et al. (1968). The characterization of amino acid sequences in proteins by statistical methods. Journal of theoretical biology https://doi.org/10.1016/0022-5193(68)90069-6

Table S1 . AAindex features which were used for creating the AAindex encoding.