Specificity of cysteine cathepsins through the eyes of large-scale proteomic data analysis

P-02.3-25

L. Tušar *[I,II], F. Impens *[III], E. Van Quickelberghe[III], M. Vizovišek[I,IV], M. Mihelič[V], R. Vidmar[I], M. Fonović[I], B. Turk[I,VI,VII], K. Gevaert[III], D. Turk[I,II]

[I]Jožef Stefan Institute, Department of Biochemistry and Molecular and Structural Biology, Ljubljana, Slovenia, [II]Centre of Excellence for Integrated Approaches in Chemistry and Biology of Proteins, Ljubljana, Slovenia, [III]VIB-UGent, Center for Medical Biotechnology and UGent Department of Biomolecular Medicine, Gent, Belgium, [IV]ETH Zürich, Inst. f. Molekulare Systembiologie, Zurich, Switzerland, [V]Mlakarjeva 11, Kamnik, Slovenia, [VI]Faculty of Chemistry, University of Ljubljana, Ljubljana, Slovenia, [VII]Institute of Regenerative Medicine, I.M. Sechenov First Moscow State Medical University, Moscow, Russia

Cysteine cathepsins are a group of endolysosomal proteases involved predominantly in protein turnover with a number of specific roles. Their redundancy and broad specificity make it difficult to understand their functional differences. To get an insight in the redundancy and specific protein degradation patterns we undertook a proteomic analysis of SH-SY5Y cell lysates with human cathepsins K, V, B, L, S, and F. Survey of approximately 30 000 cleavage sites determined by the COFRADIC method showed that the majority of cleavages can be performed by several cathepsins, with a small fraction of specific cleavages, performed by a single cathepsin in the intact protein only once. Redundant cleavages often appear in groups of short sequence spans. Statistical analysis of the cleaved peptides revealed differences in the specificity-determining substrate binding sites and their span. We call these sites heterogeneous, in contrast to homogeneous, which contain normal distribution of amino acids. The cleavage sites of heterogeneous substrate positions were clustered using the BLOSUM weighted substitution matrix and served us to extract refined training sets for generating the support vector machine models for cathepsins K, V, B, L, S, and F. We applied this methodology to successfully predict cathepsin cleavages in a number of viral proteins known to be associated with cysteine cathepsin activation, including the S-protein of SARS-CoV-2. The computational methodology developed here can be applied to provide an insight and predict interactions in a broad range of protein substrates and their processing enzymes.

* The authors marked with an asterisk equally contributed to the work.