Statistical modelling of *in vitro* pepsinolysis using peptidomic data

Ousmane SUWAREH[1], David CAUSEUR[2], Julien JARDIN[1], Valérie BRIARD-BION[1], Steven LE FEUNTEUN[1], Stéphane PEZENNEC[1], Françoise NAU[1]

1STLO, INRAE, Institut Agro, 65 rue de Saint-Brieuc, 35042 Rennes, France
2 IRMAR UMR6625, CNRS, Institut Agro, 65 rue de Saint-Brieuc, 35042 Rennes, France

E-mail for correspondence: ousmane.suwareh@agrocampus-ouest.fr

The digestion process is a complex phenomenon not yet completely understood, despite a significant amount of studies on this topic. In the case of protein foods, peptidomic data generated using mass spectrometry can be used to identify the protein fragments released due to the action of digestive enzymes, and therefore to identify the peptide bonds cleaved. Using an *in vitro* model of digestion focused on the gastric phase, our goal is to propose a statistical framework for the probability that pepsin cleaves a sequence of amino acid residues at a given peptide bond. The tested variables include the composition of the sequence itself around the peptide bond, and a large set of physicochemical features of its three-dimensional environment. The statistical framework introduces large-dimensional propensity scores, one for each amino acid residue, at each position flanking the peptide bonds, in a logistic regression model.

In order to assess the specificity of pepsin action, namely that cleavage sites along protein sequences are not distributed randomly, significance tests were first implemented. However, the large dimension of the model questioned the accuracy of the standard likelihood-ratio tests. An alternative penalized estimation procedure was also proposed to select the features favouring cleavage by pepsin, assuming that the number of amino acid residues influencing pepsin action is low. The presentation will focus on the comparison of these two approaches to analyse experimental data in a large design involving six proteins intentionally chosen to cover a large scope of physicochemical properties.

Keywords: Peptidomic data; Large-dimensional propensity scores; Pepsin specificity