

Titre : Modélisation de la pepsinolyse *in vitro* en conditions gastriques et inférence de réseaux de filiation de peptides à partir de données de peptidomique.

Mots clés : Pepsine, Spécificité, Intégration de données de peptidomique, Modèle additif généralisé, Expectation-Maximisation, Données mal étiquetées.

Résumé : Pour faire face aux enjeux démographiques actuels, aux « maladies de civilisation » et à la possible raréfaction des ressources alimentaires, il est impératif d'optimiser l'utilisation effective des aliments et d'adapter leur conception aux besoins spécifiques des différentes populations. Cela demande d'accroître notre compréhension des différentes étapes de la digestion. En particulier, en raison du rôle majeur des protéines dans notre alimentation, leur devenir au cours de la digestion est au cœur des préoccupations. Or, les lois probabilistes qui régissent l'action de la pepsine, première protéase à agir dans le tractus gastro-intestinal, ne sont pas encore clairement identifiées.

Dans une première approche s'appuyant sur des données de peptidomique, nous démontrons que l'hydrolyse par

la pepsine d'une liaison peptidique dépend de la nature des résidus d'acides aminés dans son large voisinage, mais aussi de variables physicochimiques et de structure décrivant son environnement. Nous proposons dans un second temps, tenant compte de l'environnement physicochimique à l'échelle de séquences peptidiques, un modèle non-paramétrique de l'hydrolyse de ces séquences par la pepsine et un algorithme d'estimation de type Expectation-Maximisation, offrant des perspectives de valorisation des données de peptidomique. Dans cette approche dynamique, nous intégrons les réseaux de filiation des peptides dans la procédure d'estimation, ce qui conduit à un modèle plus parcimonieux et plus pertinent au regard des interprétations biologiques.

Title-: Modelling *in vitro* pepsinolysis in gastric conditions and inferring peptide filiation networks from peptidomics data.

Keywords : Pepsin, Specificity, Integration of peptidomic data, Generalized additive model, Expectation-Maximization, Mislabeled data

Abstract: Addressing the current demographic challenges, "civilization diseases" and the possible depletion of food resources, require optimization of food utilization and adapting their conception to the specific needs of each target population. This requires a better understanding of the different stages of the digestion process. In particular, how proteins are hydrolyzed is a major issue, due to their crucial role in human nutrition. However, the probabilistic laws governing the action of pepsin, the first protease to act in the gastrointestinal tract, are still unclear.

In a first approach based on peptidomic data, we demonstrate that the hydrolysis by pepsin of a peptide

bond depends on the nature of the amino acid residues in its large neighborhood, but also on physicochemical and structural variables describing its environment. In a second step, and considering the physicochemical environment at the peptide level, we propose a non-parametric model of the hydrolysis by pepsin of these peptides, and an Expectation-Maximization type estimation algorithm, offering novel perspectives for the valorization of peptidomic data. In this dynamic approach, we integrate the peptide kinship network into the estimation procedure, which leads to a more parsimonious model that is also more relevant regarding biological interpretations.