Considering symmetries in clustering methods:
Application to the study of biomolecules

Project Supervisors:
Cathy MAUGIS-RABUSSEAU
IMT-INSA, Toulouse
cathy.maugis@insa-toulouse.fr
Juan CORTES
LAAS-CNRS, Toulouse
juan.cortes@laas.fr

Summary:

Clustering is a basic technique in statistical data analysis, and it is applied in various fields. Diverse clustering methods have been proposed over the years (HAC, k-means, DBSCAN, …), each of them having advantages and drawbacks, thus being more or less suitable for each particular problem. Moreover, in many cases, the structure of input data requires a specific treatment. This is for instance the case when data involves some symmetry.

The goal of this project is to investigate several approaches for a better treatment of symmetry within clustering methods. We will first investigate *ad-hoc* approaches to adapt basic clustering methods such as HAC to data involving symmetries known *a priori.* Then, we will investigate general approaches to automatically detect symmetries and to consider them for clustering using more sophisticated methods such as density-based or model-based clustering methods. Finally, the performance of the different methods will be evaluated and compared.

The methods developed during the project will be applied in the field of molecular modeling. More precisely, we will integrate them within computational methods to predict the structure of biomolecules on surfaces [1]. In this context, exploration and optimization methods produce molecular configurations involving symmetries induced by the crystallographic structure of the surface. These symmetries must be considered for a correct identification of configuration clusters.

References:

[1] S. Abb, N. Tarrat, J. Cortés, B. Andriyevsky, L. Harnau, J. C. Schön, S. Rauschenbach, K. Kern, Carbohydrate Self□Assembly at Surfaces: STM Imaging of Sucrose Conformation and Ordering on Cu(100). *Angewandte Chemie*, 131, 8424, 2019,

Expected skills:

Strong background in statistics is mandatory, as well as good programing skills (Python, C++, R). Background in structural biology is not necessary, but it would be a plus.

Possibility of funding:

The student will be provided with a monthly stipend of around 550 euros during up to six months.

Applications:

Please send an email containing your CV to Cathy Maugis-Rabusseau (cathy.maugis@insa-toulouse.fr) and Juan Cortés (juan.cortes@laas.fr), indicating in the subject “Candidate clustering project”.