Integrative Modeling of Viral Proteomic Features for Predicting Host Specificity in Plant Viruses

Nikolay Simankov [1,2,*], Rachid Tahzima [1], Hélène Soyeurt [2], Sébastien Massart [1]

[1]Laboratory of Plant Pathology – TERRA – Gembloux Agro-BioTech – University of Liège (ULiège) – 5030 Gembloux, Belgium.

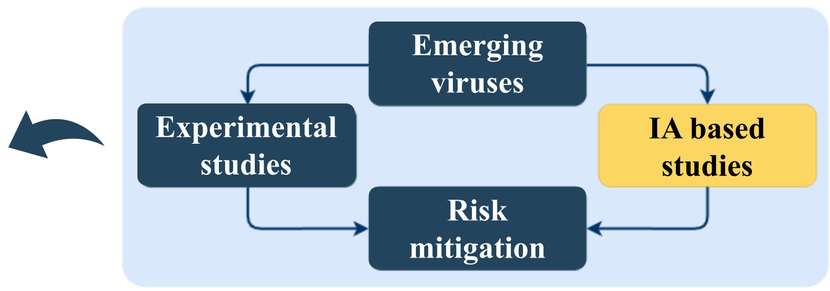
[2]Statistics, Computer Science and Modeling applied to bioengineering – TERRA – Gembloux Agro-BioTech – University of Liège (ULiège) – 5030 Gembloux, Belgium.

*This Communication is supported by the Fond National de la Recherche Scientifique (FNRS) as part of a FRIA grant.

Plant Viruses

Plant viruses constitute a vast and phylogenetically diverse group of pathogens posing significant threats to global agricultural productivity and food security. They represent a **major epidemiological challenge** due to their rapid dissemination and complex transmission dynamics. Indeed, viruses account for over **50% of emerging diseases** affecting cultivated plants, often resulting in substantial economic losses, reaching billions **of euros annually**.

In recent years, the advent of **high-throughput sequencing** technologies has dramatically increased the volume of genomic data, with thousands of new viral genomes added monthly to public repositories such as the Sequence Read Archive (SRA). However, **biological characterization** of these newly discovered viruses often remains **incomplete**, with approximately 20% of known plant virus species currently lacking comprehensive biological information, focusing solely on genomic characterization. This underscores the **critical need for integrative studies that combine genomic data with biological insights to effectively manage and mitigate viral threats in agriculture**.

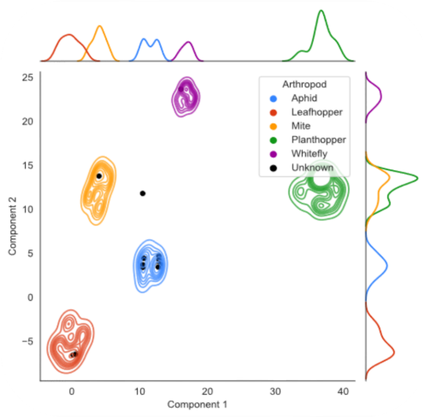


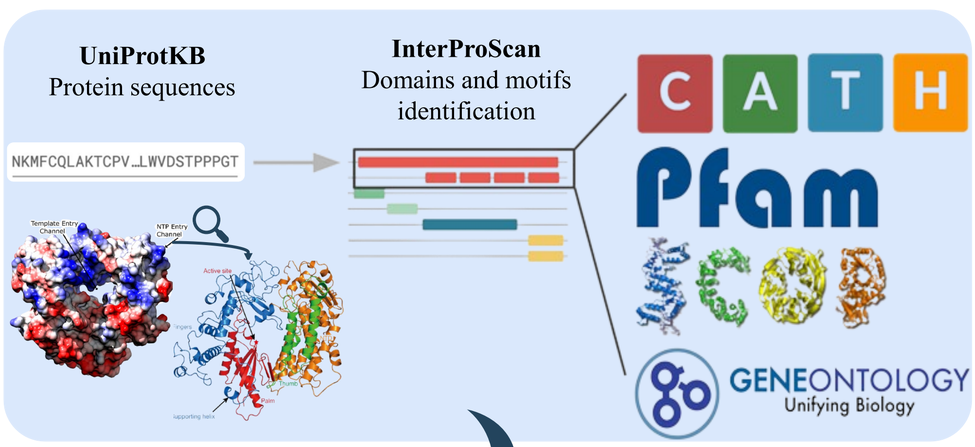
Testing relationships experimentally

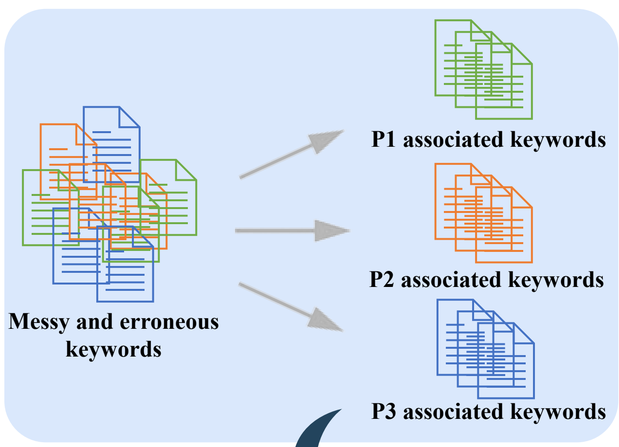
- Not exhaustive
- Time-consuming
- Expensive
- Based on observations
- Sometimes inconclusive

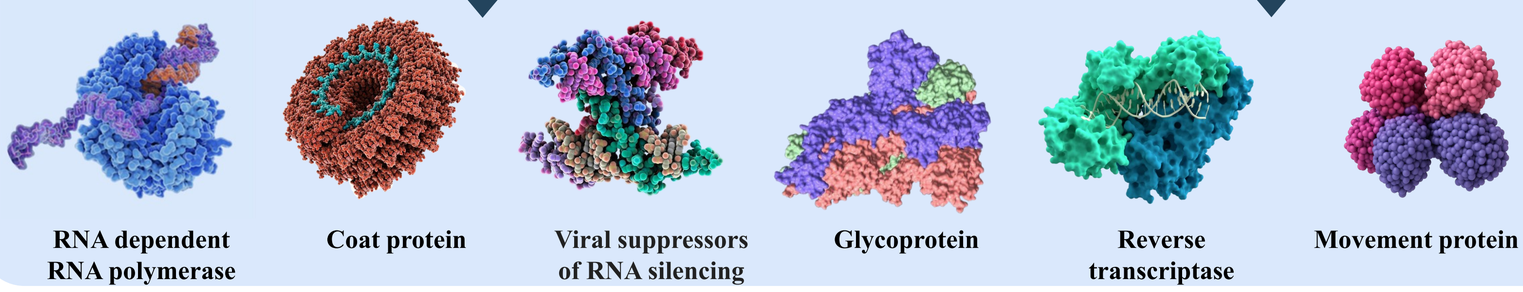
Predicting hosts by proteomics features

- Allows quick processing of large amount of data
- Based on complete genomes (HTS)
- Uses high-performance bioinformatics tools
- Helps understanding of virus/host interactions









Proteomic Data and Associated Challenges

The rapid **data accumulation** in major **biological databases** such as UniProt, EMBL, and GenBank (NCBI) has created **unprecedented opportunities** for comparative genomics and functional proteomics research. Nonetheless, in Virus Kingdom, the **absence of standardized protein naming conventions** across different taxa considerably complicates their effective integration, annotation accuracy, and comparative analyses, ultimately limiting the full potential of these extensive biological resources.

Using a comprehensive dataset comprising over **150,000 viral protein sequences** from **UniProtKB**, we successfully standardized the nomenclature of common viral proteins across the entire viral kingdom. This was achieved through an integrated methodology combining the identification of **structural and functional motifs** with advanced text analysis and **synonym clustering** techniques. Such an approach dramatically improves the consistency of protein naming and **functional annotation** conventions, facilitating **comparative analyses across taxa**, unlocking the full analytical potential of proteomic datasets for virus research.

Linking Physico-chemical Protein Properties to the Biological Context With HARAMO

We generated about **9,000 physicochemical features** for each protein using the **ProtLearn** pipeline. A database of more than **28,000** known plant host-virus **relationships** was constituted.

Using a Holistic AutoML-driven Robust pipeline optimization tool for Applied Multi-Omics (**HARAMO**) that we designed, we identified key physicochemical signatures (**amino acid composition, properties of secondary structures and intrinsic disorder**) of proteins that are involved in **virus-host specificity** in plants. Our **protein-based approach achieved MCC scores in host plant prediction ranging from 79.6 to 98.6%**, depending on the input viral protein and the target plant, offering predictive insights for epidemiological surveillance of emerging threats.

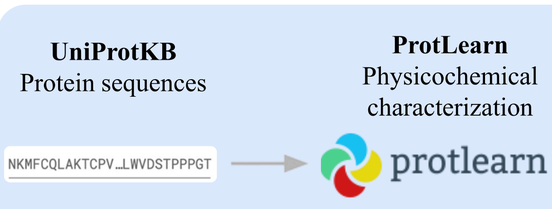
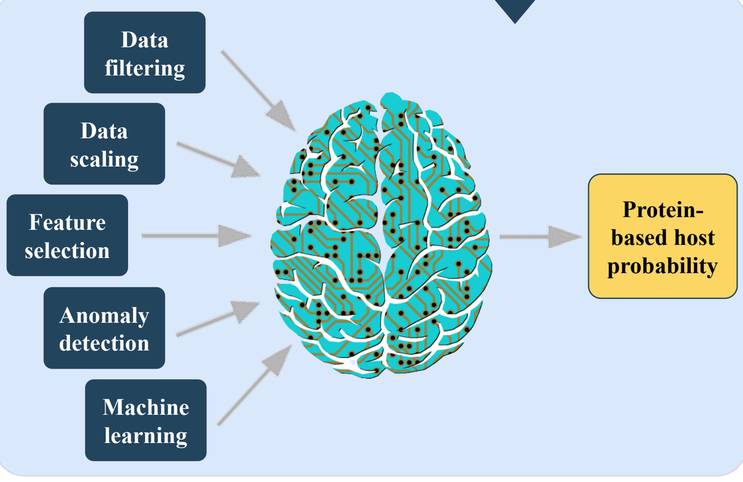


Table 1: Physicochemical properties

Descriptor groups	Descriptor	Number of Features
Amino Acid Composition	*Amino acid composition*	20
	Dipeptide composition	400
	Tripeptide composition	8000
Conjoint Triads (CTD)	*Conjoint Triad - Composition*	21
	Conjoint Triad - Transition	21
	Conjoint Triad - Distribution	105
Conjoint Triads	*Conjoint Triads*	343
Intrinsic Disorder (LCider)	*Amino acid properties*	25
	Total	**8935**



HARAMO

Haramo is a **two-step** python **AutoML** package conceived to automatically design a complete **pipeline** adapted to the data before optimizing hyperparameters. Moreover, it is especially suited to process **multi-omics** data and to distinguishing noisy data from informative data.

ex : Potyvirus yituberosi

Virus/Host relationship by PredOmics

Overall, we have developed and trained over 5,000 protein-based prediction pipelines to explore complex virus–host relationships. These models are integrated into an interactive online **dashboard** that enables predictions based on both our protein database and user-submitted **FASTA/FASTQ** protein sequences. This tool is designed to support **experimental research**, **enhance epidemiological surveillance**, and assist decision-making by relevant authorities.

Our integrative framework offers a robust resource to better understand virus–host dynamics and will serve as a foundation for biological validation in collaboration with a newly formed **consortium of partner laboratories** specializing in plant health.

Eventually, this tool may be extended to animal viruses to meet public health needs.