Internship on structure, co-evolution and design of pore-forming antimicrobial peptides

Environment :

- **Team**: DAO and GruLab, LJK CNRS, https://grulab.imag.fr/ + DPM, COMET team
- **Working place**: IMAG, Grenoble Alpes University, 38401 St Martin d'Hères (Grenoble)
- **Research topics**: deep learning, bioinformatics, generative models

Introduction: A promising strategy to fight against antimicrobial resistance can be designing novel antimicrobial peptides (AMPs). These are short amino acid sequences (typically ranging from 6 to 50 residues in length) that kill various bacteria, viruses and fungi mainly through membrane disruption, but also a few other mechanisms[1,2,3]. Recently, protein design[4,5] (proteins are long peptides, longer than 50 residues) has been transformed by AI tools such as RF-Diffusion, ProteinMPNN, and AlphaFold (awarded the 2024 Nobel Prize in Chemistry)[6,7,8]. Building on these advances, this project aims to transfer the protein-specific AI developments to peptides by reusing structural and sequence patterns in proteins with tandem repeats[9].

Methods: Our main source of structural and sequence information will be RepeatsDB[10] and EncoMPASS[11] databases. RepeatsDB is a repository for the annotation and classification of structural tandem repeat proteins (STRPs). Each entry has start and end positions of the repeat region, repeat units, classification into four levels (Class, Topology, Fold, Clan), and in-depth characterization of the repeat regions. RepeatsDB contains 5,138 closed tandem-repeated proteins and 21,030 AF2 predicted models. It also contains 2,809 elongated repeats and 5,543 corresponding AF2 models. Our working hypothesis is to use available co-evolution and structural information about tandem-repeated proteins to redesign individual repeat subunits in form of peptides with the hope that they will self-assemble together in vivo driven by non-covalent interactions. The disadvantage of this database that it only contains 375 membrane proteins.

RepeatsDB uses the definition of structured tandem repeat proteins, which is characterized by a minimum of 3 structural repeat units. In the current version, STRPs are identified using the STRPsearch tool. Consequently, repeats composed of only 2 units and very short repeats (e.g., homorepeats, dipeptide repeats) are not included. If we check for reviewed transmembrane proteins in the UniProt using the "repeat" feature filter, we get total of 2,410 proteins. Thus, running other repeat detection tools could expand the RepeatsDB dataset. EncoMPASS can be another option that seem to have more repeated fragments. However, their definition is different from those used in UniProt and RepeatsDB.

In the first part of the project, we will annotate repeated proteins that penetrate the membrane cell and constitute the training set and the training-test splits. ***This part is very important and may take significant amount of time.*** Then, we will use a cyclicMPNN version of proteinMPNN to redesign their sequences conditioned by the structure and membrane contacts. For this reason, we will specifically construct and train a membrane cyclic-MPNN version of the protein design architecture using the constructed training set. Successful designs will be tested in-silico by AlphaFold refold ability scores and then synthesized and tested in cell and liposome models by the DPM partner.

Requirements: We are looking for creative, passionate and hard-working individuals from applied math / computer science background with exceptional talent for computer science and mathematics and interest in computational biology / chemistry / physics. Excellent oral, written and interpersonal communication skills are essential (working language will be English – knowledge of French is a plus). Good knowledge of PyTorch / C++ / signal processing / machine learning / structural bioinformatics will be an asset.

Advisors: Sergei Grudinin (LJK, sergei.grudinin@univ-grenoble-alpes.fr), Clovis Galiez (LJK, clovis.galiez@univ-grenoble-alpes.fr), and Yung-Sing Wong (DPM, yung-sing.wong@univ-grenoble-alpes.fr).

Keywords: generative models; deep learning; protein design, geometric deep learning

References:

1. Szymczak, Paulina, et al. "AI-Driven Antimicrobial Peptide Discovery: Mining and Generation." Accounts of Chemical Research (2025): 1453-871.
2. Lai, Leshan, et al. "Deep generative models for therapeutic peptide discovery: A comprehensive review." ACM Computing Surveys 57.6 (2025): 1-29.
3. Wan, Fangping, et al. "Machine learning for antimicrobial peptide identification and design." Nature Reviews Bioengineering 2.5 (2024): 392-407.
4. Albanese, K I., et al. "Computational protein design." ***Nature Reviews Methods Primers*** 5.1 (2025): 13.
5. Wang, C., et al. "Toward deep learning sequence–structure co-generation for protein design." ***Current Opinion in Structural Biology*** 91 (2025): 103018.
6. Watson et al., Nature (2023).
7. Dauparas et al., Science (2022).
8. Jumper et al., Nature (2021).
9. Kajava, AV. "Tandem repeats in proteins: from sequence to structure." J of structural biology 179.3 (2012): 279-288.
10. Clementel et al. "RepeatsDB in 2025: expanding annotations of structured tandem repeats proteins on AlphaFoldDB", Nucleic Acids Research, Volume 53, Issue D1, 6 January 2025, Pages D575–D581.
11. Aleksandrova AA, Sarti E, Forrest LR. EncoMPASS: An encyclopedia of membrane proteins analyzed by structure and symmetry. Structure. 2024 Apr 4;32(4):492-504.