Projet de Recherche Doctoral Concours IPV 2021

Intitulé du Projet de Recherche Doctoral : AIDE: Artificial Intelligent Directed Evolution

Directeur de Thèse porteur du projet (titulaire d'une HDR) :

NOM: **FERRARI** Prénom : **Ulisse**

Titre : **CRCN-CNRS**

e-mail : ulisse.ferrari@gmail.com / ulisse.ferrari@inserm.fr

Adresse professionnelle : **17 rue Moreau, 75012, Paris**

Unité de Recherche :

Intitulé : Institut de la Vision

Code : UMRS 968 -- UMR 7210 - UM 80

Equipe de Recherche *(au sein de l'unité)* **:**

Intitulé : Visual information processing: neural coding and vision restoration

Thématique de recherche : - transmission of visual information in normal or pathological conditions
- new therapeutic strategies to prevent vision loss and restore vision in blind patients

Responsable d'équipe :

NOM : **PICAUD** Prénom : **Serge**

Ecole Doctorale de rattachement de l'équipe & d'inscription du doctorant : ED3C: Cerveau, cognition, comportement

Doctorants actuellement encadrés par le directeur de thèse (préciser le nombre de doctorants, leur année de 1ere inscription et la quotité d'encadrement) :

MAHUAS Gabriel, Octobre 2020, 50%: co-encadrement avec T. Mora (LPENS)

CO-DIRECTION *(obligatoire)*

Co-Directeur de Thèse (titulaire d'une HDR) :

NOM : **DALKARA** Prénom : **Deniz**

Titre : **INSERM-DR** HDR Oui

e-mail : deniz.dalkara@inserm.fr

Unité de Recherche :

Intitulé : Institut de la Vision

Code: UMRS 968 -- UMR 7210 - UM 80

Equipe de Recherche *(au sein de l'unité)* **:**

Intitulé : Gene therapies and animal models for neurodegenerative diseases

Thématique de recherche : Approaches of gene delivery to develop effective gene-based therapies

Responsable d'équipe :

NOM : **DALKARA** Prénom : **Deniz**

Ecole Doctorale de rattachement : Physiologie, Physiopathologie et Thérapeutique

Doctorants actuellement encadrés par le co-directeur de thèse (préciser le nombre de doctorants, leur année de 1ere inscription et la quotité d'encadrement) :

- Müge Tekinsoy, 100%, February 2017
- Catherine Botto, 100%, January 2019
- Marco Rucli, 100%, March 2020

Cotutelle internationale : Non Oui, précisez Pays et Université : **NON**

Résumé (2 000 caractères maximum) :

Directed evolution is a versatile strategy for engineering proteins like enzymes, antibodies or viral vectors. Recently this experimental technique started to benefit from the massive sequence analysis provided by deep-sequencing techniques. Such massive data call for the integration of computational approaches into directed evolution to further empower this engineering strategy.

In this thesis proposal, the candidate will focus on the massive deep-sequencing data collected from two unique experiments where directed evolution has been applied to engineering novel viral vectors for more efficient gene delivery to the human retina. His/her goal will be to develop and infer machine learning models in order to learn the directed evolution process, and use them for engineering empowered viral vectors. Compared to mostly-unpublished analyses previously attempted on these data, a machine learning approach will allow us to deal better with the high noise and the high dimensionality of the data.

This is a very interdisciplinary project between biology and machine learning, where state-of-the-art computational techniques are applied to groundbreaking biological experiments. The candidate will benefit from the experimental data collected over multiple rounds of protein screening on two different species, and uses them to train machine-learning models to predict the fitness of the different viral variants along the directed evolution workflow. To accomplish this goal, the candidate will have the possibility to interact with, and learn from researchers from both a biological and a computational team.

Joindre en annexe un descriptif du PRD avec références au format pdf (« NOM_2_IPV_2021 » / 3 pages maximum, taille police 11)

AVIS et VALIDATION de l'ECOLE DOCTORALE :

Voir l'avis sur la page ajoutée en fin de document

ECOLE DOCTORALE 3C
CERVEAU COGNITION COMPORTEMENT
Sorbonne Université
Université de Paris
PSL
Directeur : Prof. Alain Trembleau

à envoyer simultanément par e-mail à l'ED de rattachement et au programme : ***interfaces_pour_le_vivant@listes.upmc.fr avant le lundi 15 février minuit.***

AIDE - Artificial Intelligent Directed Evolution

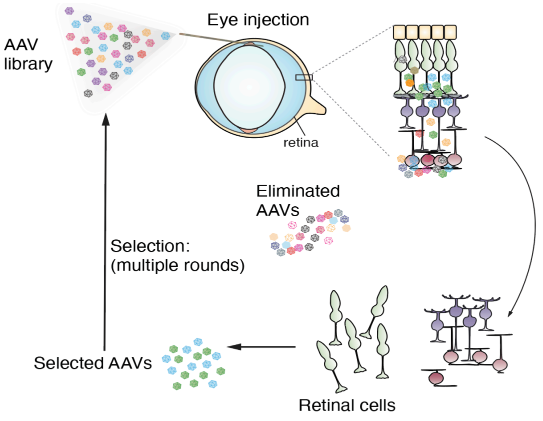

Fig. 1: DE enriches the variants capable of transfecting the retina.

After the FDA approval of Luxturna to treat Leber congenital amaurosis, great progress has been made with retinal gene therapy using adeno-associated viral (AAV) vectors. However, **a therapeutically impactful transduction of retinal neurons is still problematic, especially in humans** (Bennett et al., 2017): limited biological knowledge obstructs a rational design of such AAV variants. To fill this unmet medical need, **directed evolution (DE) has been applied to engineer AAV variants that transfect the human retina, but clearcut results are still lacking.** DE (Arnold 1998, 2018's Chemistry Nobel prize) is an engineering strategy for proteins like enzymes (Chen et al. 1993), antibodies (Boder et al 2000), or viral vectors (Dalkara et al. 2013). DE emulates natural evolution (Fig. 1): it starts from billions of random variants, it iteratively screens them against a given task and finally unveils the best performing variants.

Today, DE benefits from the tremendous progress in DNA sequencing techniques called Next Generation Sequencing (NGS) or deep sequencing. NGS allows for reading of millions of distinct DNA sequences at each round of the DE screening iterations. These massive data call for the integration of machine learning approaches to improve DE performance (Bedbrook et al., 2019). Our goal is **to upgrade the results of two complementary DE experiments by inferring computational models from these massive NGS data**. The goal is to learn a model of the DE screening process to overcome high dimensionality and data-poor limitations, to use the model to predict high performing variants, and eventually to patent an efficient viral vector for human retinas.

State of the art and preliminary results

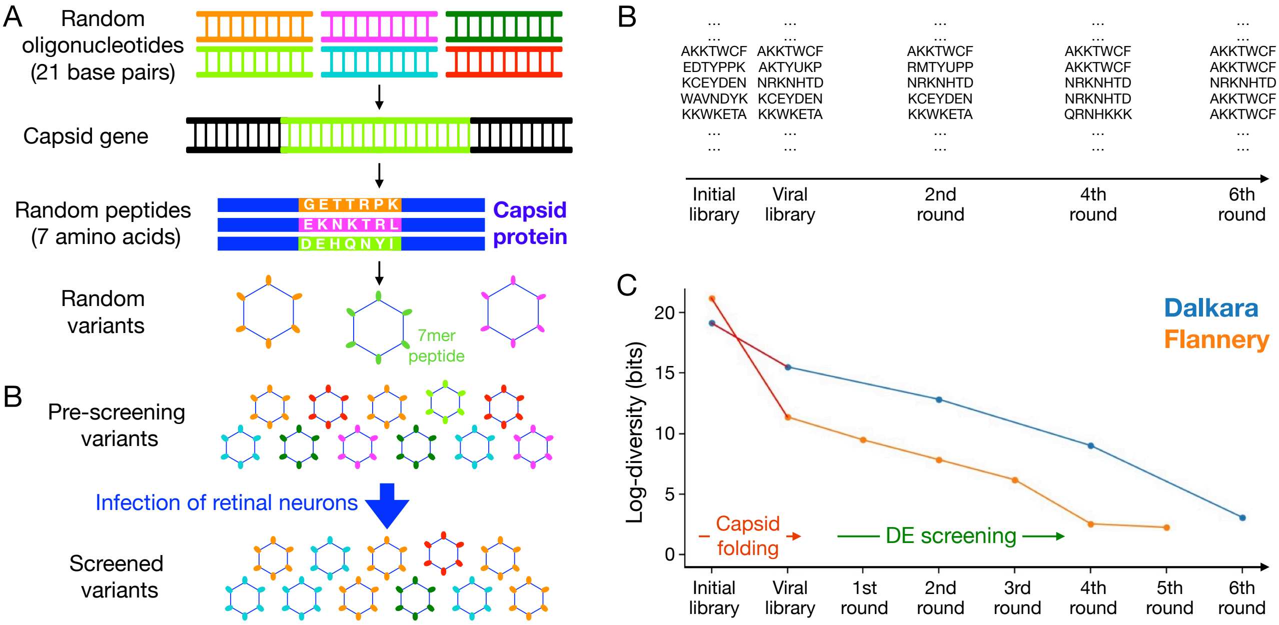

Fig. 2: DE experiment on AAV vectors. A) Billions of different random oligonucleotides are inserted in the wildtype viral genome (Initial library). Synthesis and capsid folding result in a Viral library of random variants. B) DE screens variants by selecting those that can infect retinal neurons (Fig. 1) C) Deep NGS allows for reading of millions of variants' DNA at multiple rounds. D) DE screening shrinks down the library log-diversity (entropy, estimated from the NGS data) along the experimental rounds up to few bits, thus unveiling few 'evolved' variants.

DE for engineering AAV vectors works on four steps (fig.s 1&2). First, a highly diverse initial library is generated by inserting 7 amino-acid random peptides in the AAV capsid protein (Fig. 2A). Peptide insertions allow for improved transduction by altering receptor binding, cell entry and intracellular trafficking. Then, multiple rounds of DE screening enrich the high performing variants (Fig. 2B). Here the screening task consists in selecting the variants capable of infecting the retinal neurons (Fig. 1). Later, massive NGS allows for reading millions of variant DNAs in the library at each step (fig. 2C), and to follow the convergence onto few 'evolved' variants (fig. 2D).

We previously used *in-vivo* DE to engineer an AAV capable of delivering genetic material into cells of the **mouse retina** (Fig. 1, Dalkara et al., 2013). This DE study unveiled "AAV2-7m8", an AAV variant which has been immensely valuable for gene delivery to the mouse retina (Mace et al. 2015; Khabou et al. 2018). Yet, **there are tremendous interspecies differences** between the performance of viral vectors (Planul & Dalkara 2017): **AAV2-7m8, which has been developed on mice, is suboptimal for transducing retinas of large animals like humans**.

In order to develop an efficient AAV vector for humans, research had started focusing on larger animals. At Flannery's lab. in Berkeley University, *In-vivo* DE studies of AAV on canine retinas (Byrne et al., submitted, ***dataset-1***) and non-human primate retinas (Byrne et al. 2020) identified viable vectors for the corresponding species, which however are suboptimal for humans**.** To focus directly on humans, in our lab, we run a DE experiment screening on post-mortem **human tissue samples** (Planul et al, in preparation, ***dataset-2***). Candidate variants have been identified, but the results on human retinal explants were not satisfactory (only a few fold increase in transduction was obtained). **The PhD work will mainly focus on datasets-1 and 2.**

All these DE experiments on large animals were able to converge onto few 'evolved' variants (fig. 2D), showcasing the screening power of the approach. However the final performance of selected variants were unsatisfying. *If the DE experiment worked as expected, why was it not able to identify a sufficiently strong variant?* Preliminary results show that **the reason lies in how the experimental outcome was analysed.** The presence of over-represented sequences at the beginning (Fig. 3), result in an unfair competition between variants: excellent variants can not overcome good ones, if the latter are far more abundant at the beginning. Consequently, **looking at the variants to which DE converges is not the right strategy.** Studying the enrichment of individual variants is neither a viable solution because the number of possible variants is orders of magnitude larger than data size, resulting in a dimensional crisis where bulk noise hinders even more the enrichment of those unfairly penalised variants. To experimentally overcome both these issues many more DE rounds would have been required, well beyond feasibility. Because of its excessive costs, reducing the bulk-noise by increasing the number of NGS reads is neither a viable solution. **We therefore need a more efficient analysis to identify the best variants. As such, we propose to learn ML models of the whole DE screening process.**

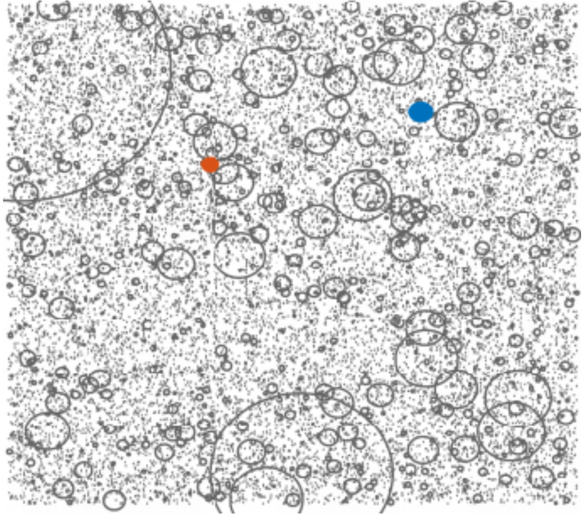

Fig. 3: the viral library before DE: each circle is a variant whose radius is proportional to its concentration. 'Evolved' variants (blue and red dots), have midrange initial concentration between few overrepresented and many underrepresented variants.

Work program

Preliminary results show that because of the heterogeneity of the viral library and the large bulk noise, **many promising variants were not able to emerge, despite being present in the sample**. We therefore need an approach that encompasses single variants and their noise by analysing the behaviour of the whole experiment altogether. For this, instead of analysing the enrichment of one individual sequence at the time, we will infer from data models of the whole DE screening. **We will use ML models capable of decomposing the sequences into amino-acid motifs and studying their contribution to the overall sequence enrichment.**

Working on DE experiments on enzymes, Fernandez-de-Cossio-Diaz et al. (2020) has proved that it is possible to combine data from multiple DE rounds and multiple sequences to learn a statistical model capable of predicting the enrichment of individual variants across successive DE rounds. Specifically, their pairwise Potts model takes as input the amino-acid sequence, decomposes it into all the possible position-dependent pairs of amino-acids, learns the synergy of all couples through the model's pairwise interactions and eventually predicts the change in concentration upon a DE iteration. Importantly, during model-learning, and thanks to this pairwise decomposition, the model overcomes the noise of individual variants by grouping the information from all the variants that share the same pair of amino-acids. **We aim at applying a similar strategy to our dataset, but tailored to our case.** Preliminary results have shown that pairs of amino-acids play a major role in determining the variant enrichment in our DE process, but also that triplets and higher order motifs cannot be ignored. To account for this important difference, **we will consider more flexible models than pairwise Potts, as for example non-linear latent models,** which have already showcased their efficiency for deep sequencing data (Riesselman et al. 2018), **or Restricted Boltzmann Machines,** which have been proved very powerful in accounting for protein statistical properties (Tubiana et al. 2019).

We aim at hiring a PhD student that will work on models' development, inference and benchmark (Months: 1-20). Once an effective model for the variant enrichment will be inferred from data, we will use it to identify the amino-acid sequence(s) with the highest predicted enrichment (Months:20-26). This candidate virus will then be tested thanks to the facilities of Dalkara's lab. both *in-vitro,* on a *post-mortem* human retina sample, and *in-vivo,* through intravitreal injection in a macaque eye. Obtaining a strong GFP expression **will allow us to showcase the power of the viral vector, patent its sequence, and largely accelerate the development of genetic therapies for vision restoration.**

Team

U. Ferrari is a tenured CNRS researcher (sect. 51, System Biology) at the Vision Institute with a long-standing experience in data-analysis and machine learning applied to neurosciences (Ferrari et al 2020, Mahuas et al 2021, Sorochynskyi et al 2021). After a PhD in statistical physics with G. Parisi in Rome, he did a first postdoc at ENS-Paris and a second one at the Vision Institute.

D. Dalkara is an INSERM research director at the Vision Institute and has broad expertise in gene delivery vectors (Dalkara et al., 2013), retinal gene therapy and retinal dystrophies. After a PhD in biology in Strasbourg (awarded the Biovalley PhD thesis) she did a first postdoc at the Ernst Babmerg's lab. at the Max Planck Institute of Biophysics, and then a second one at UC Berkeley (Euretina Science and Medicine Innovation award in 2013 and selected Innovator under 35 –France by MIT Technology Review in 2014).

The team joins partners with complementary and interdisciplinary expertise in both DE for virus engineering and ML applied to biological data. They have already worked together on vision restoration using gene therapy (Ferrari et al. 2020) and currently jointly supervise a senior postdoc, a M2 student (our potential candidate) and, informally, an experimental PhD. This project will strengthen their collaboration and prime a rich new set of interdisciplinary interactions based on Dr Dalkara's other ongoing DE experiments.

Feasibility

The stake of this project is very high because a new viral vector for gene delivery to the human retina would be extremely valuable for clinical gene therapy applications for fighting blindness. Yet the risks are moderate because we already have all the experimental data we need.

The project is in continuity with the just-expired ERC starting grant of Dr Dalkara (2014) and will synergistically complement its experimental outcome. A complementary project has already been founded through our institute's IHU program (1.5 years, mostly covering the postdoc salary). In that proposal, we focus on the development of AI and molecular biology tools for the generation of a smart initial library that reduces the biases in the viral library discussed above.

References

- Arnold, F. H. (1998). *Accounts of chemical research*, *31*(3), 125-131.
- Bedbrook, C. N., Yang, K. K., ..., Arnold, F. H. (2019). *Nature methods*, 1-9.
- Bennett, A., Patel, S.,…, Agbandje-McKenna, M. (2017). *Mol. Ther.-Meth & Clin. D.*, *6*, 171-182.
- **Byrne L.C., …, Dalkara D.,…,Flannery J.G. (2020) JCI Insights, 5, 10 Open**
- Boder, E. T., Midelfort, K. S., Wittrup, K. D. (2000). *PNAS*, *97*(20), 10701.
- Chen, K., & Arnold, F. H. (1993). *PNAS*, *90*(12), 5618-5622.
- **Dalkara, D., Byrne, …, Schaffer, D.V. (2013). *Sci. Transl. Med.*, *5*(189), 189ra76.**
- Fernandez-de-Cossio-Diaz, J., Uguzzoni, G., Pagnani, A. (2021) Mol. Biol. evol. 38 (1) **Open**
- **Ferrari, U.,…, Dalkara, D, et al. (*2020) PLOS Comput. Biol. 16 (7), e1007857.* Open**
- **Khabou, H., Cordeau, ..., Dalkara, D. (2018). *Human gene therapy*, *29*(11), 1235-1241.**
- **Macé, E., Caplette, …., Dalkara D. (2015). *Molecular Therapy*, *23*(1), 7-16.**
- **Mahuas G.,…, Ferrari* U, Mora* T (2021). NeurIPS-2021 Open**
- **Planul, A., Dalkara, D. (2017). *Annual review of vision science*, *3*, 121-140.**
- Riesselman, A.J., Ingraham, J.B., Marks, D.S. (2018) *Nature methods*,*15*(10), 816-822.
- **Sorochynskyi O, …, Marre* O., Ferrari* U. (2021) *PLOS Comput. Biol. 17(1), e1008501* Open**
- Tubiana, J., Cocco, S., & Monasson, R. (2019). *Elife*, *8*, e39397. **Open**