Projet OBAMA:
Augmentation des données basée sur l'orthologie

raphael.mourad@univ-tlse3.fr

Raphaël Mourad, Julie Demars, Céline Brouard

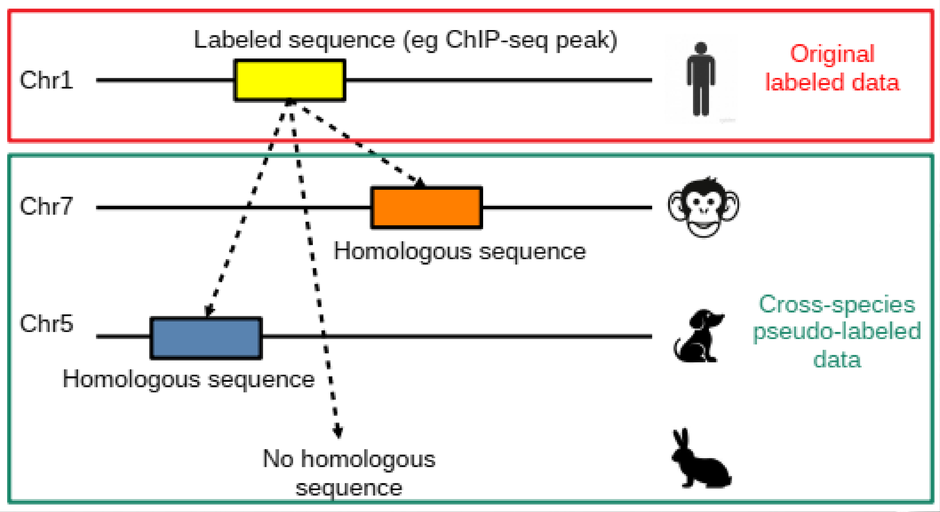
MIAT INRAE, GenPhyse INRAE, Université Paul Sabatier

Introduction

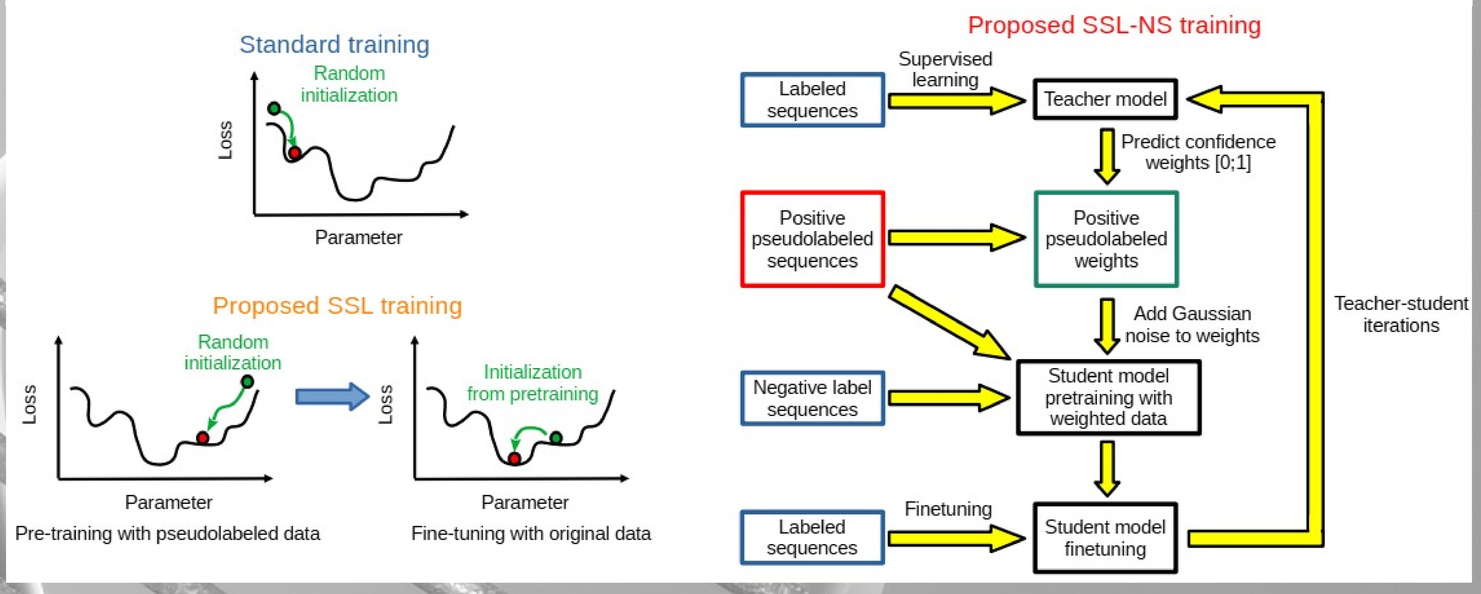
La caractérisation génomique et fonctionnelle des animaux d'élevage apparait comme un levier pour la transition agroécologique via, entre autres, l'identification des liens génotype-phénotype. Les études d'association pangénome ont identifié des milliers de variants associés à des caractères agronomiques complexes. Cependant, la majorité de ces variants ont été trouvés dans des régions génomiques non codantes, empêchant la compréhension du mécanisme biologique sous-jacent.

Prédire les processus moléculaires basés sur la séquence d'ADN à l'aide des méthodes par apprentissage profond représentent une approche prometteuse pour comprendre le rôle de ces variants non codants. L'apprentissage classique, supervisé, nécessite des séquences d'ADN associées à des données fonctionnelles pour l'entraînement, dont la quantité est fortement limitée par la taille finie du génome humain. Cependant, les approches d'augmentation des données par orthologie permettraient d'enrichir considérablement les jeux de données d'entraînement et améliorer ainsi la capacité prédictive des modèles.

Augmentation de données annotées par orthologie



Algorithmes d'apprentissage supervisé (SL) et semisupervisé (SSL)



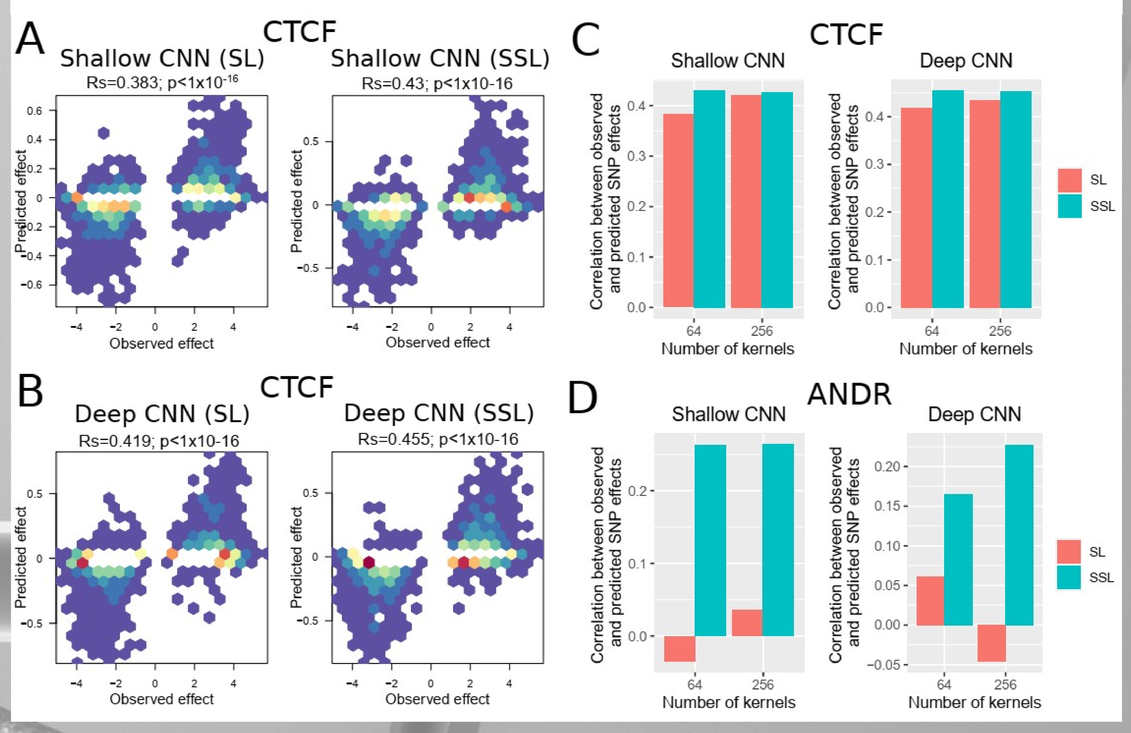
Taille des données avant et après augmentation

Data	Number of positive sequences	Number of pseudo-labeled Positive sequences	Increase
ATF3	1306	23555	18x
ETS1	4211	77661	18x
ANDR	4638	36164	8x
REST	7201	130615	18x
MAX	23916	441295	18x
P300	15844	272923	17x
RAD21	37113	646495	17x
CTCF	75082	1245195	17x
H3K4me3	67589	1220489	18x
POL2	57226	998107	17x
ATAC	70600	1186950	17x

Benchmark et comparaison avec DNABERT2

		AUPR					
Data	Number Of peaks	Simple CNN 7K parameters Supervised Learning	Simple CNN 7K parameters SSL	Simple CNN 27K parameters Supervised Learning	Simple CNN 27K parameters SSL	Simple CNN 27K parameters SSL-NS	DNABERT2 117M parameters
ATF3	1677	0.008	0.124	0.021	0.149	**0.176**	0.039
ETS1	4120	0.035	0.103	0.030	0.139	**0.183**	0.052
ANDR	4638	0.002	0.0041	0.002	**0.0042**	0.003	0.002
REST	6119	0.044	0.204	0.062	**0.253**	0.246	0.015
MAX	13605	0.077	0.104	0.087	0.1100	0.109	**0.1103**
P300	14223	0.006	0.030	0.012	**0.043**	0.036	0.012
RAD21	34623	0.283	0.303	0.257	**0.331**	0.324	0.172
CTCF	72779	0.404	0.461	0.414	0.4835	**0.4844**	0.308
H3K4me3	25641	0.128	0.156	0.154	0.165	0.170	**0.188**
POL2	35982	0.118	0.143	0.132	0.159	**0.166**	0.131
ATAC	102030	0.083	0.109	0.086	0.119	**0.121**	0.112
#1st place		0	0	0	4	**5**	2

Prediction de l'impact de SNP



Conclusion

Dans cet article, nous avons proposé un nouvel apprentissage semi-supervisé (SSL) basé sur des données pseudo-étiquetées permettant d'entraîner des modèles à partir de données ayant plusieurs ordres de grandeur par rapport aux données étiquetées disponibles. Nous l'avons encore amélioré en intégrant les principes de l'algorithme Noisy Student pour prédire la confiance dans les données pseudo-étiquetées.

De telles approches atténuent la limite de taille des génomes humains (3,3 Go), en exploitant d'autres génomes de mammifères. Étant donné que la quantité de génomes de mammifères séquencés augmente de façon exponentielle en raison de projets de séquençage à grande échelle (par exemple le projet Zoonomia), de telles approches représentent une voie prometteuse pour améliorer les performances des modèles d'apprentissage profond actuels.

References:

- Han Phan, Céline Brouard, Raphaël Mourad*. Pre-training with pseudo-labeling compares favorably with large language models for regulatory sequence prediction. Briefings in Bioinformatics, 2024.