

Создание инструмента для аннотации регуляторных вариантов в геноме человека

Научный руководитель – Пензар Дмитрий Дмитриевич

Мурашка Вероника Васильевна

Студент (специалист)

Московский государственный университет имени М.В.Ломоносова, Факультет
биоинженерии и биоинформатики, Москва, Россия

E-mail: myra2003@yandex.com

По текущим оценкам, порядка 80-90% геномных вариантов человека, ассоциированных с клинически значимыми заболеваниями, приходится на регуляторные участки генома. На данный момент наиболее применяемым инструментом для аннотации VCF файлов является CADD. При составлении аннотации VCF файла он полагается на разнообразные биологические параметры: метрики консервативности, расстояние до ближайших экзон-интронных границ, уровень экспрессии и др. Для предсказания эффектов регуляторных мутаций в CADD 1.6 были добавлены подходы на основе искусственных нейронных сетей - в его основе лежит конволюционная нейронная сеть [3]. Однако на соревновании DREAM Challenge 2022 CADD показали довольно скромные результаты, заняв только 20 место, не превзойдя по качеству даже базовое решение, предложенное организаторами конкурса [2]. В этом же соревновании была предложена архитектура LegNet, которая стала победителем во всех категориях [1].

Помимо этого, существует несколько различных вариантов аннотации подобных вариантов, каждый из которых обладает своими плюсами и недостатками. В CADD реализован лишь один, выбранный авторами.

На фоне этого было решено создать новый инструмент для автоматической аннотации VCF файлов на основе LegNet. Были предобучены версии модели для разных тканей и клеточных линий. Полученный инструмент был протестирован на базах данных ADAstra и UDACHA, а также на имеющихся данных о регуляторных мутациях из базы данных ClinVar.

Источники и литература

- 1) Dmitry Penzar, Daria Nogina, Elizaveta Noskova, Arsenii Zinkevich, Georgy Meshcheryakov, Andrey Lando, Abdul Muntakim Rafi, Carl de Boer, Ivan V Kulakovskiy, LegNet: a best-in-class deep learning model for short DNA regulatory regions, Bioinformatics, Volume 39, Issue 8, August 2023, btad457, <https://doi.org/10.1093/bioinformatics/btad457>
- 2) Rafi, A.M., Nogina, D., Penzar, D. et al. A community effort to optimize sequence-based deep learning models of gene regulation. Nat Biotechnol (2024). <https://doi.org/10.1038/s41587-024-02414-w>
- 3) Rentzsch P, Schubach M, Shendure J, Kircher M. CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Med. 2021 Feb 22. doi: 10.1186/s13073-021-00835-9

Иллюстрации

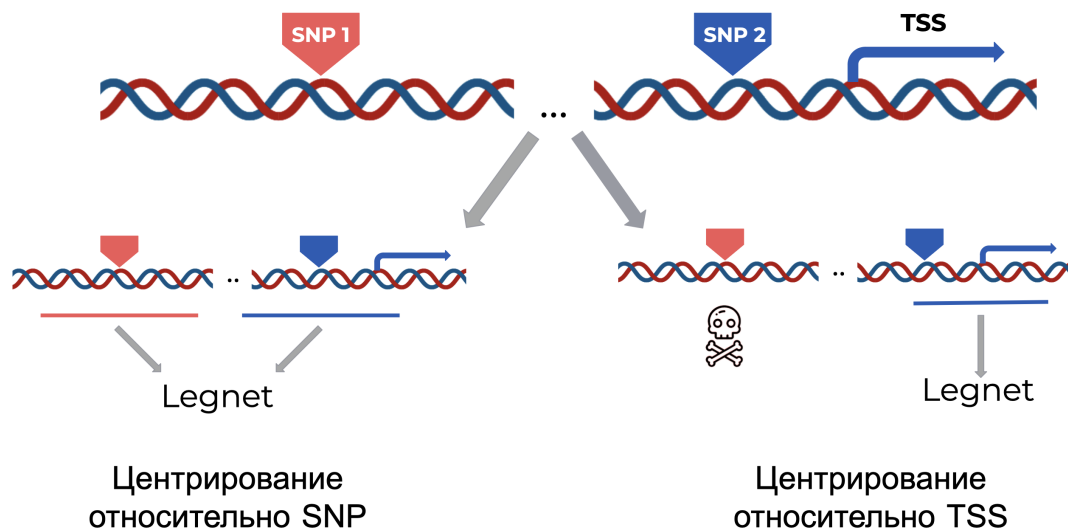


Рис. : Оценка эффекта SNP возможна в двух вариантах: в окне, центрированном относительно мутации, и в окне, картированном относительно TSS.

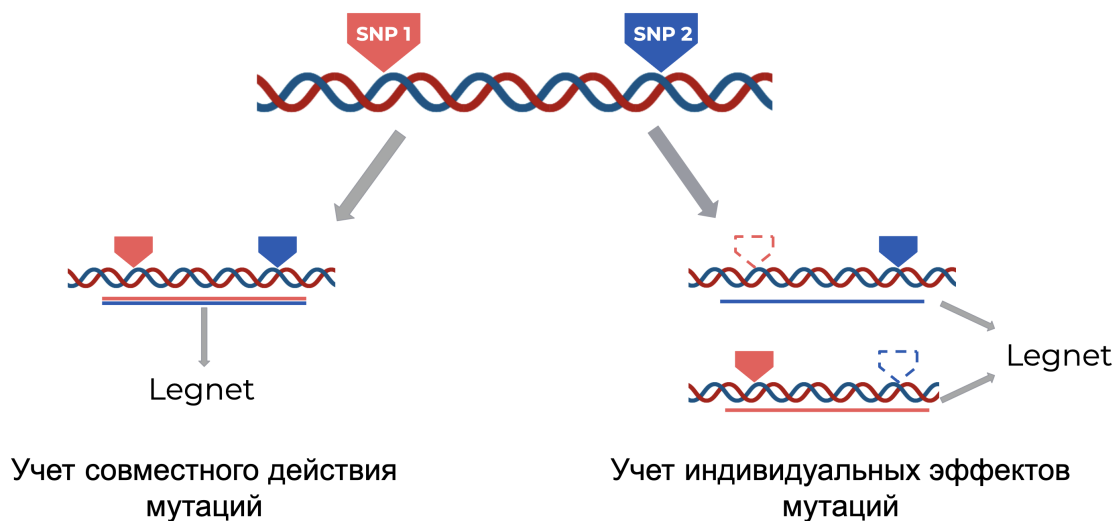


Рис. : Вторая реализованная опция заключается в том, следует ли учитывать все мутации в заданном окне или же сосредоточиться на одной конкретной. Это позволит рассматривать как и взаимодействие целого ансамбля мутаций, так и сделать предсказания индивидуальных SNP.