Секция «Биоинформатика»

Использование данных о связывании факторов транскрипции для предсказания регуляторных эффектов однонуклеотидных вариантов

Научный руководитель – Кулаковский Иван Владимирович

Зинкевич Арсений Олегович

Студент (специалист)

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

E-mail: arseniilog@list.ru

Предсказание влияния однонуклеотидных замен (SNV, single nucleotide variants) в регуляторных районах на экспрессию генов является важной биоинформатической задачей, решение которой необходимо для персонализированной диагностики и лечения заболеваний, имеющих генетическую компоненту.

В ходе участия в открытом соревновании по машинному обучению CAGI 2018 "Regulation Saturation" в работе (*Пензар и др., 2018*) мы предложили новый вычислительный метод для предсказания эффекта вариантов на основе данных высокопроизводительных экспериментов с репортерами (massively parallel reporter assays, MRPA). Наш метод опирается в первую очередь на геномную последовательность регуляторного района и результаты ее обработки специализированной нейросетью DeepSEA [1], которые служат признаками для обучения классификатора "случайный лес".

Для обучения и валидации использовались данные о изменении экспрессии для всех SNV в 9 промоторах (F9, GP1BB, HBB, HBG, HNF4A, LDLR, MSMB, PKLR, TERT) и 5 энхансерах (IRF4, IRF6, MYC, SORT1, ZFAND3). В качестве дополнительной валидации использовались данные по двум энхансерам (ALDOB, ECR11), опубликованные ранее [2].

Мы изучили вклад дополнительных независимых признаков: эпигенетических разметок (карт связывания факторов транскрипции и доступности хроматина) и эволюционной консервативности.

В качестве исходных данных для эпигеномных признаков были использованы: база данных Gene Transcription Regulation Database [3][4], данные ENCODE (ATAC-Seq и DNase-Seq), FANTOM5 CAGE [5]. Консервативность последовательности оценивалась по данным из UCSC Genome Browser [6].

Использование эпигеномных признаков и информации о консервативности не позволило достигнуть качества предсказаний, основанных на DeepSEA-признаках. Тем не менее, использование полученных признаков в ансамбле с нейросетевыми повысило качество предсказания.

Источники и литература

1) Predicting the Effects of Noncoding Variants with Deep learning-based Sequence Model. Jian Zhou, Olga G. Troyanskaya. Nature Methods (2015)

2) Massively parallel functional dissection of mammalian enhancers in vivo. Rupali P Patwardhan et al. Nat Biotechnol. 2012 Mar; 30(3): 265–270.

3) GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. I.S. Yevshin, R.N. Sharipov, T.F. Valeev, A.E. Kel, F.A. Kolpakov. Nucleic Acids Res. 2017 Jan 4;45(D1):D61-D67.

4) GTRD: a database on gene transcription regulation—2019 update. I.S. Yevshin, R.N. Sharipov. S.K. Kolmykov, Y.V. Kondrakhin, F.A. Kolpakov. Nucleic Acids Res. 2018 Nov 16; gky1128

5) Gateways to the FANTOM5 promoter level mammalian expression atlas. Lizio M, et al. Genome Biol 16: 22 (2015). 10.1186/s13059-014-0560-6

6) Detection of non-neutral substitution rates on mammalian phylogenies. Pollard KS, Hubisz MJ, Siepel A. Genome Res. 2010 Jan;20(1):110-21.