Санкт-Петербургский государственный университет

СОСНИН Юрий Константинович

Выпускная квалификационная работа

Алгоритмы валидации вариантных пептидов

Образовательная программа бакалавриат «Математика»
Направление и код: 01.03.01 «Математика»
Шифр ОП: СВ.5000.2019

Научный руководитель:
д.ф.-м.н., доцент
Факультет МКН СПбГУ
Степанов Алексей Владимирович

Рецензент:
к.ф.-м.н., Научный сотрудник,
Институт энергетических проблем
химической физики им. В.Л. Тальрозе
ФГБУН Федерального исследовательского
центра химической физики
РАН им. Н.Н. Семенова
Иванов Марк Витальевич

Санкт-Петербург
2023 год

Содержание

1 Введение

Пептиды, содержащие аминокислотные полиморфизмы (мутации), называются вариантыми. Задача детекции вариантых пептидов актульна тем, что данные полиморфизмы вызывают нарушения в работе белка, что может привести к развитию различных заболеваний.

В экспериментах в протеомике на основе жидкой хроматографии в тандеме с масс-спектрометрией (liquid chromatography-tandem mass spectrometry, LC-MS/MS) пептиды предварительно разделяются при помощи жидкой хроматографии и потом передаются масс-спектрометру. Время, за которое хроматограф подаёт пептид спектрометру называется временем удерживания (retention time, RT). Так как время удерживания основывается на физических и химических свойствах пептида при определённом виде жидкой хроматографии, оно предсказуемо в теории [1].

Целью данной работы является рассмотрение возможности применения времени удерживания как признака пептида для детекции вариантных пептидов.

2 Введение в предметную область

В настоящий момент все самые распространённые методы идентификации пептидов так или иначе основываются на масс-спектрометрии. Одними из таких методов являются поиск и сопоставление по базам данных, сопоставление по спектрам, ДНК- и РНК-секвенирование, de novo секвенирование, идентификация по отпечаткам масс (mass fingerprinting) или же комбинации этих методов [2]. Схему применения этих подходов для обнаружения вариантных пептидов [3] можно видеть на рис. 1. Здесь рассматривается SAP - single amino acid polymorphism, то есть вариантный пептид с одной мутацией. Как можно видеть, одна эта мутация сразу меняет показания масс-спектрометра.

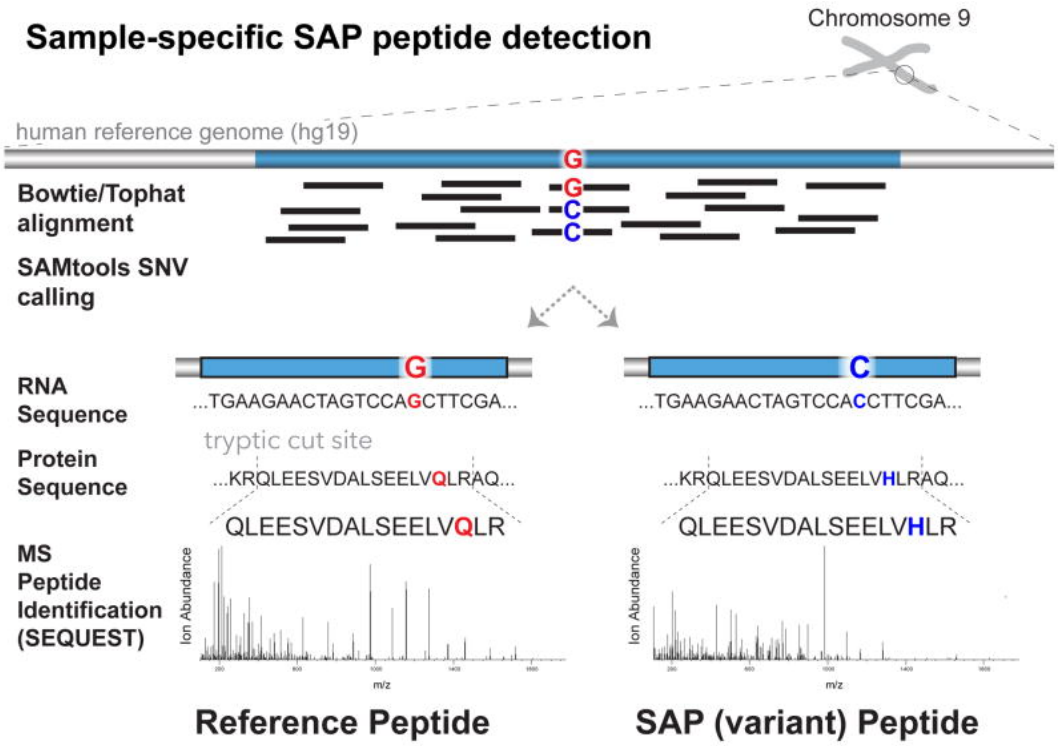

Рис. 1: Схема идентификации вариантного пептида

В данный момент основным подходом для идентификации вариантных пептидов является подключение двух видов баз данных к результатам масс-спектрометрии: общедоступные базы данных уже встречавшихся вариантных пептидов и базы данных вариантных пептидов, полученных при помощи ДНК- и РНК-секвенирования [4]. Для поиска по базам данных используются такие программы как Sequest, Mascot, MyriMatch и X!Tandem. Но у обоих этих подходов есть свои недостатки, в первом случае имеет место существенная вероятность ошибочного обнаружения пептида из-за слишком большого размера самой базы, во втором же случае, наоборот, может произойти необнаружение последовательности из-за вычислительной ошибки на стадии секвенирования [5].

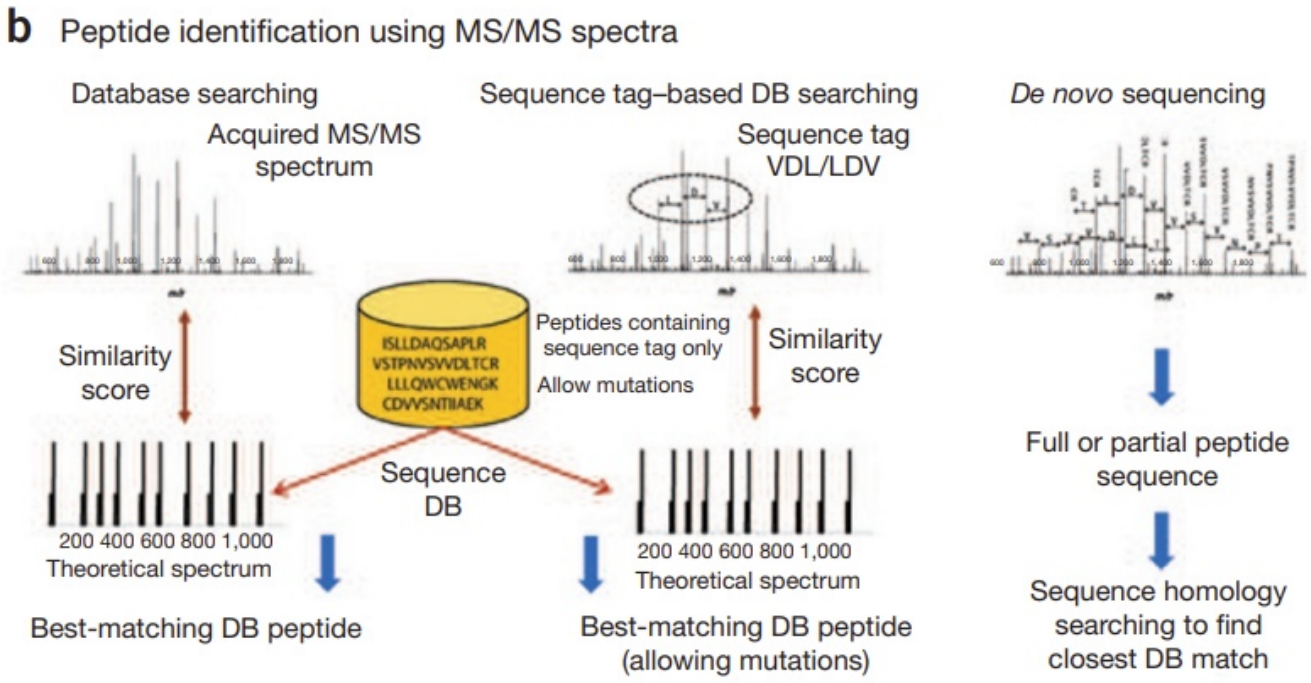

Рис. 2: Методы идентификации [2]

Конечно, рост популярности нейросетей и алгоритмов машинного обучения не мог обойти и протеомику. В последние года было разработано программное обеспечение на основе глубокого обучения не только для предсказания времени удерживания, но и для непосредственной идентификации пептидов по масс-спектрам [6], хотя последнее оставлено за рамками данной работы. Мы же сконцентрируемся на различных моделях, основанных на глубоком и машинном обучении для предсказания времени удерживания пептидов.

3 Рассматриваемое программное обеспечение

3.1 DeepRT(+)

DeepRT был представлен в работе [7] и использует глубокое обучение, а именно свёрточные нейронные сети (Convolutional Neural Network, CNN) для извлечения признаков пептидов, что исключает необходимость делать это вручную. После извлечения признаков DeepRT использует три различных метода машинного обучения для непосредственного предсказания времени удерживания на основе выделенных признаков: метод опорных векторов (support vector machine, SVM), случайный лес деревьев (random forest, RF) и градиентный бустинг (gradient boosting, GB). И финальный результат выдаётся при помощи бэггинга трёх моделей (рис. 3).

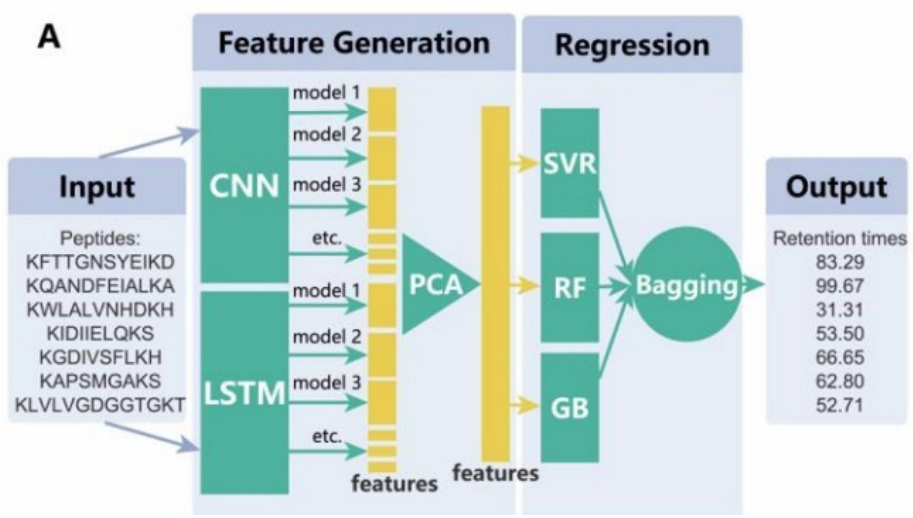

Рис. 3: Архитектура DeepRT [8]

3.2 DeepLC

DeepLC [9] тоже использует свёрточные нейронные сети, но с немного другим подходом (рис. 4). Каждый пептид подаётся на вход четырём путям: one-hot encoding, извлечение глобальных признаков (длина пептида, масса), аминокислотный состав и диаминокислотный состав. Во всех путях, кроме извлечения глобальных признаков, используются конволюционные и подвыборочные (max pooling) слои. В оставшемся пути используются плотно связанные слои (densely connected layers). Результаты всех четырёх путей уплощаются, конкатинируются и подаются на вход финальному пути, состоящему из шести плотно связанных слоёв.

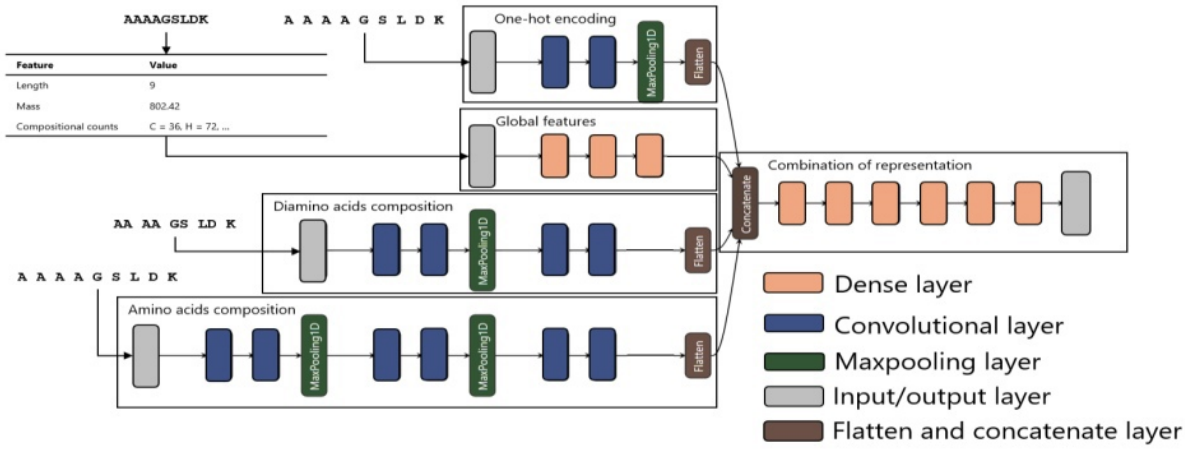

Рис. 4: Архитектура DeepLC[3]

3.3 Pyteomics.achrom

pyteomics.achrom является самой базовой библиотекой языка программирования Python для предсказания времени удерживания. Она основывается на аддитивной модели и выдаёт результат исходя лишь из аминокислотного состава и длины пептида. Время удерживания предсказывается исходя из следующей формулы:

$$RT = (1 + m \ln N) \sum_{i=1}^{N} RC_i n_i + RT_0$$

где m - поправочный коэффициент длины, N - длина пептида, RC_i - коэффициент задержки аминокислоты типа i, n_i - количество аминокислот типа i, RT_0 - константный сдвиг времени удерживания. Здесь, RC_i находятся во время калибровки (обучения) модели.

Наименование датасета	Размер трейна	Размер теста	Кол-во уникальных последовательностей на трейне	Количество уникальных последовательностей на тесте
HeLa	70986	7887	29882	6171
HeLa с EV<0.1	41448	4606	11918	3329
Наш датасет	4900	548	1147	308
Наш датасет с EV<0.1	1111	123	139	53

Таблица 1: Сводная статистика датасетов

4 Рассматриваемые датасеты

В нашей работе мы использовали два датасета.

В качестве первого датасета был взят датасет HeLa, состоящий из 78873 последовательностей, взятых из раковых клеток. Здесь наблюдаемое время удерживания составляло от 9 до 240 минут.

Второй датасет был экспериментальным и имел довольно "грязные"данные. Всего использовалось 5448 пептидов, но среди них было достаточно много повторений последовательностей, много "недорезов"(к примеру, встречались такие последовательности как AVVQDPALKPLALVYGEATSR и AVVQDPALKPLALVYGEATSRR), а также само качество данных, оцениваемое при помощи EValue, далеко не всегда было приемлемым (часто EValue превышало 1). Наблюдаемое время удерживания у данных последовательностей варьировалось от 0,3 до 40 минут.

Отличительной чертой обоих этих датасетов стало присутствие аминокислотных модификаций, а именно карбомидометилирования цистеина (Cys). В исходных mzXML файлах данная модификация отмечается как C+57.021.

5 Результаты

Все датасеты предварительно были перемешаны и разбиты на датасет для обучения и датасет для теста в отношении 9:1. Для оценки качества работы моделей в качестве метрики был выбран коэффициент корелляции Пирсона r:

$$r = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}}$$

Для начала было принято решение проверить работу данных моделей на широко известном датасете - HeLa. В данном датасете получилось 7887 последовательностей на тесте и 70986 на обучении. На всех графиках по горизонтали отложено реальное время удерживания, по вертикали предсказанное время удерживания. Шкала для датасета HeLa от 0 до 240 минут.

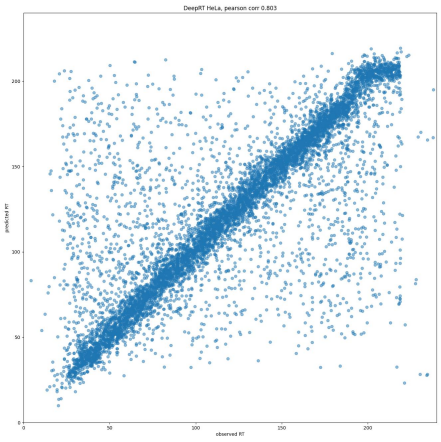

Рис. 5: HeLa на DeepRT+

Рис. 6: HeLa на DeepLC

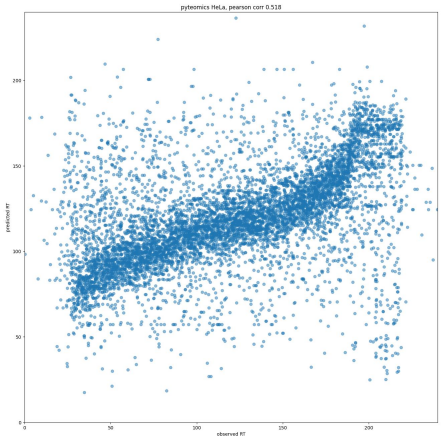

Рис. 7: HeLa на Pyteomics

На данном датасете лучше всего себя проявил DeepRT+ (рис. 5), коэффициент корелляции Пирсона для данной модели составила 0.803, прослеживается чёткая диагональ, хоть и с большим шумом. DeepLC (рис. 6) дал странные результаты, все предсказания данной модели лежат в промежутке от 60 до 190 минут, а корелляция Пирсона равна 0.626. Pyteomics.achrom (рис. 7) же вместо желаемой диагонали дал что-то больше похожее на кубическую кривую и с сильным разбросом и корреляцией Пирсона 0.518.

Для всех последовательностей из датасетов было известно их EValue - метрика качества скана с масс-спектрометра или же степень уверенности для данной последовательности. Чем меньше EValue, тем качественней данный скан. Для дальнейшей работы было принято выбрать из датасета HeLa только последовательности с EValue меньше 0.1. Новый датасет содержал 4606 последовательностей на тесте и 41448 последовательностей на обучении.

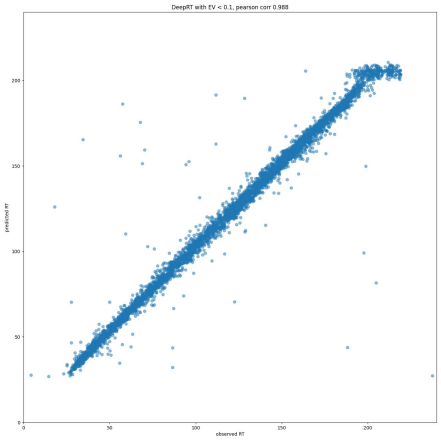

Рис. 8: HeLa EV<0.1 на DeepRT+

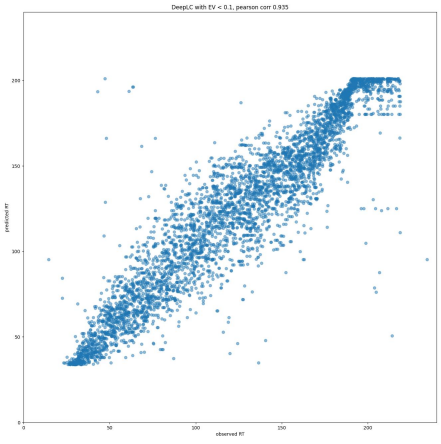

Рис. 9: HeLa EV<0.1 на DeepLC

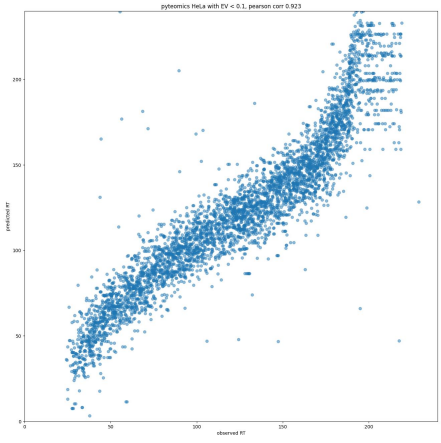

Рис. 10: HeLa EV<0.1 на Pyteomics

Все три модели показали себя гораздо лучше, но лучше всего опять был DeepRT+ (рис. 8) с очень явной диагональю, малым разбросом и $r = 0.988$. DeepLC (рис. 9) опять предсказал значения только из подмножества реальных значений, но на этот раз от 40 до 200, а Pyteomics.achrom (рис. 10) дал ещё более явную кубическую кривую.

Далее данные модели были протестированы на нашем экспериментальном датасете. На тесте получилось 543 последовательности, но из-за его особенностей различных из них оказалось всего 231. Ниже представлены графики рассеивания для предсказаний трёх моделей на нашем датасете.

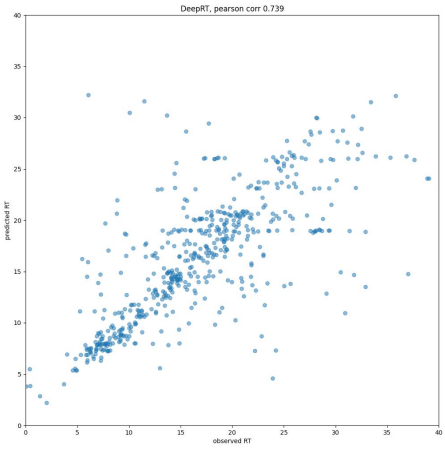

Рис. 11: Наш датасет на DeepRT+

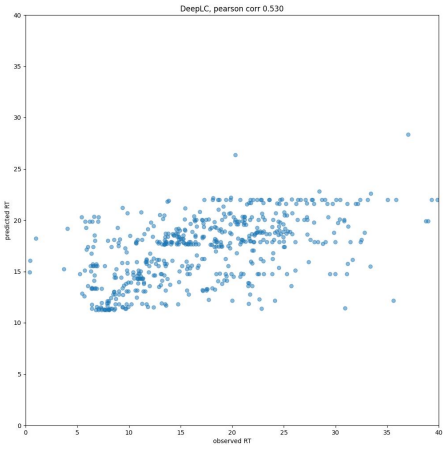

Рис. 12: Наш датасет на DeepLC

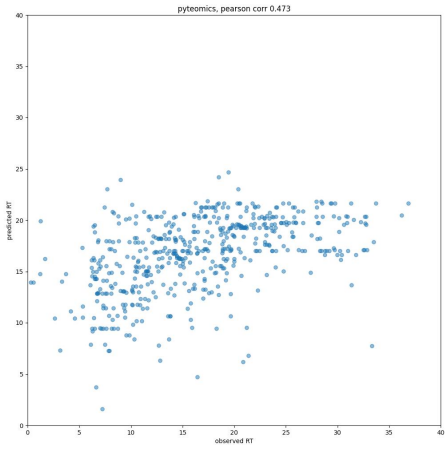

Рис. 13: Наш датасет на Pyteomics

Здесь же приемлемые результаты показал только DeepRT+ (рис. 11), диагональ прослеживается, хоть и разброс уже очень сильный (корреляция Пирсона в данном эксперимен-

те составила около 0.739). Но вот DeepLC (рис. 12) сработал с такими же данными плохо. Во-первых, абсолютно не прослеживается желаемой диагонали, $r = 0.53$, во-вторых, за исключением редких выбросов, все предсказания сосредоточены на отрезке между 10 и 20 минутами, что совсем не отражает реальной картины. Похожая картина и для аддитивной модели pyteomics.achrom (рис. 13) - отсутствует видимая диагональ, корреляция $r = 0.473$ и все предсказания лежат на отрезке от 5 до 23.

Последним датасетом для рассмотрения стало подмножество этого датасета с EValue $<$ 0.1. Здесь количество последовательностей на тесте стало уже 123, а на трейне 1111.

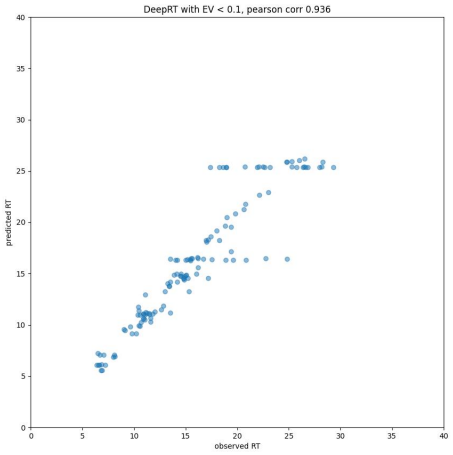

Рис. 14: Наш датасет с EV$<$0.1 на DeepRT+

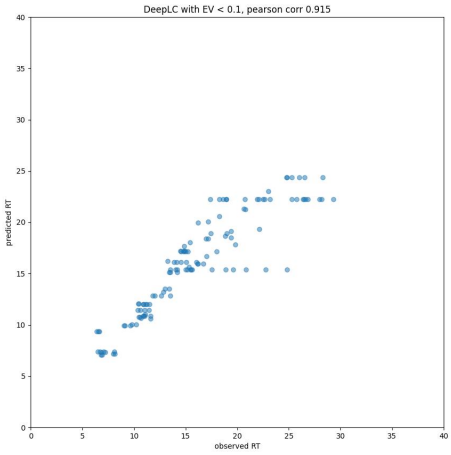

Рис. 15: Наш датасет с EV$<$0.1 на DeepLC

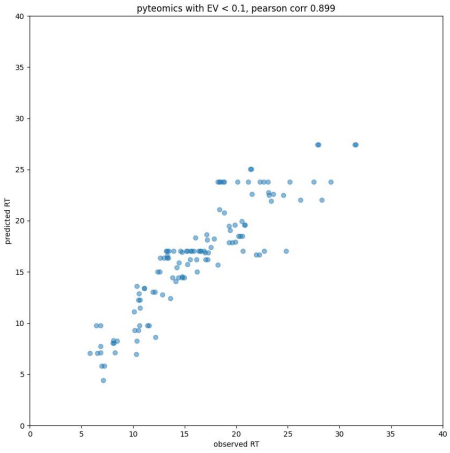

Рис. 16: Наш датасет с EV$<$0.1 на Pyteomics

В этом случае все три модели проявили себя хорошо, даже аддитивная модель Pyteomics.achrom (рис. 16) получила предсказания с $r = 0.899$. Лучшей моделью опять был

DeepRT+ (рис. 14) с $r = 0.936$, но и DeepLC (рис. 15) был рядом с корелляцией $r = 0.915$. Отличительной особенностью результатов данного эксперимента стали явные горизонтальные отрезки в графиках предсказаний. После более близкого изучения результатов стало ясно, что эти отрезки представляют собой предсказания для одинаковых последовательностей, как, например, на рисунке 17. Здесь колонки слева направо: последовательность, EValue, реальное время удерживания, предсказанное время удерживания DeepRT+, абсолютная разница. Как можно видеть, для данной последовательности реальное время удерживания варьировалось от 17 до 28 минут, а DeepRT+ дал предсказание в 25.3 минут, отсюда и появляется горизонтальный отрезок на графике.

AVVQDPALKPLALVYGEATSR	4,01E-16	22,483	25,3988	2,9158
AVVQDPALKPLALVYGEATSR	2,17E-13	28,17167	25,3988	2,772867
AVVQDPALKPLALVYGEATSR	1,25E-14	26,45667	25,3988	1,057867
AVVQDPALKPLALVYGEATSR	9,63E-17	22,11333	25,3976	3,284267
AVVQDPALKPLALVYGEATSR	4,21E-17	20,76483	25,3968	4,631967
AVVQDPALKPLALVYGEATSR	5,07E-13	25,27483	25,3968	0,121967
AVVQDPALKPLALVYGEATSR	2,62E-15	17,37417	25,3524	7,978233
AVVQDPALKPLALVYGEATSR	8,23E-13	18,96717	25,3524	6,385233
AVVQDPALKPLALVYGEATSR	1,07E-13	27,9805	25,3524	2,6281
AVVQDPALKPLALVYGEATSR	2,71E-16	26,63467	25,3524	1,282267
AVVQDPALKPLALVYGEATSR	7,47E-12	25,75233	25,3524	0,399933
AVVQDPALKPLALVYGEATSR	2,88E-11	29,334	25,352	3,982
AVVQDPALKPLALVYGEATSR	2,45E-17	22,66167	25,352	2,690333
AVVQDPALKPLALVYGEATSR	3,56E-15	23,149	25,352	2,203
AVVQDPALKPLALVYGEATSR	1,32E-14	26,81483	25,352	1,462833
AVVQDPALKPLALVYGEATSR	4,32E-21	18,28267	25,3452	7,062533
AVVQDPALKPLALVYGEATSR	1,06E-18	18,64017	25,3452	6,705033
AVVQDPALKPLALVYGEATSR	1,61E-21	18,923	25,3452	6,4222
AVVQDPALKPLALVYGEATSR	2,22E-16	21,93333	25,3452	3,411867
AVVQDPALKPLALVYGEATSR	6,76E-13	26,36783	25,3452	1,022633

Рис. 17: Фрагмент нашего датасета

6 Заключение

В данной работе было исследовано устройство самых передовых моделей для предсказывания времени удерживания пептидов, проведены эксперименты для данных моделей как на широкоизвестных, так и на экспериментальных данных и построены соответствующие графики рассеивания с вычисленными коэффициентами корелляции Пирсона. Модель DeepRT+ проявила себя лучше всего, и, в теории, успешные предсказания должны помочь повысить качество идентификации вариантных пептидов в комбинации с уже известными методами. В дальнейшей работе планируется поближе изучить последовательности, в которых данные модели ошибаются и предложить методы по их улучшению, а так же выявить, какие признаки могут помочь успешно "почистить"датасет для лучшего качества предсказаний помимо EValue.

Список литературы

[1] Moruz, L. and Käll, L. *Peptide retention time prediction.* Mass Spectrometry Reviews 2016:n/a-n/a.

[2] Nesvizhskii, A. I. *Proteogenomics: concepts, applications and computational strategies.* Nature Methods 11, 1114–1125. issn: 1548-7091 (2014).

[3] Gloria M. Sheynkman, Michael R. Shortreed, Brian L. Frey, Mark Scalf, and Lloyd M. Smith *Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences* Journal of Proteome Research 2014 13 (1), 228-240

[4] Menschaert, G. and Fenyö, D. *Proteogenomics from a bioinformatics angle: A growing field. Mass Spec Rev, 36: 584-599.* https://doi.org/10.1002/mas.21483 (2017)

[5] Alioto TS, Buchhalter I, Derdak S, Hutter B, Eldridge MD, Hovig E, et al. *A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing.* Nat Commun. 2015;6:10001.

[6] Feng S, Sterzenbach R, Guo X. *Deep learning for peptide identification from metaproteomics datasets. J Proteomics.* 2021 Sep 15;247:104316. doi: 10.1016/j.jprot.2021.104316. Epub 2021 Jul 8. PMID: 34246788; PMCID: PMC8435027.

[7] Ma, C. et al. *Improved peptide retention time prediction in liquid chromatography through deep learning.* Anal. Chem. 90, 10881–10888 (2018)

[8] Ma, C. et al. *DeepRT: deep learning for peptide retention time prediction in proteomics.* arXiv Prepr. arXiv1705.05368 (2017).

[9] Bouwmeester, R., Gabriels, R., Hulstaert, N. et al. *DeepLC can predict retention times for peptides that carry as-yet unseen modifications.* 10.1101/2020.03.28.013003 (2020).

[10] Alfaro, J.A., Ignatchenko, A., Ignatchenko, V. et al. *Detecting protein variants by mass spectrometry: a comprehensive study in cancer cell-lines.* Genome Med 9, 62 (2017).

[11] Mason, K. E., Anex, D., Grey, T., Hart, B., Parker, G. (2018). *Protein-based forensic identification using genetically variant peptides in human bone.* Forensic Science International, 288, 89–96.