Поиск оптимальной стратегии формирования обучающих выборок для моделей эпитоп-МНС

Антон Смирнов

Mar 20, 2023

Оглавление

Введение

Для поиска оптимальной стратегии формирования выборок и уровня дескрипторов для обучения моделей эпитоп-MHC были сформированы из данных таблицы mhc_bind Immune Epitope Database и данных, используемых для обучения MHCflurry, но полученные не масс-спектрометрией, обучающие выборки. Обучающие выборки были сделаны 2 типов.

1. Комбинированная, где в один файл объединены данные об активности и неактивности эпитопов избранных 10 аллелей MHC
2. Отдельные, где данные по активности и неактивности к конкретному избранному аллелю MHC собраны в отдельные файлы

Необходимость проверки заключаются в том, что PASS формирует отрицательные примеры для класса из всех случаев, которые не принадлежат ему. Это не совсем правильно с биологической точки зрения, так как взаимодействие одного эпитопа с одним аллелем MHC не исключает, что этот же эпитоп будет взаимодействовать с другим аллелем. Поэтому второй подход биологически верен, но он может не обеспечить достаточной точности прогноза. Кроме этого необходимо установить оптимальный уровень дескрипторов, который обеспечивает максимальную точность.

Методы

Для моделирования были выбраны 5 аллелей MHC высокопредставленных в имеющихся данных и 5 аллелей имющих низко- или среднюю представленность.

Table 1: Избранные аллели MHC

Высокопредставленные	Низкопредставленные
HLA-A*02:01	HLA-A*02:16
HLA-A*03:01	HLA-C*14:02
HLA-A*11:01	HLA-C*15:02
HLA-A*02:03	HLA-C*04:01
HLA-A*02:06	HLA-B*73:01

Представленность аллелей

activity	unique_epi
!HLA-A*02:01	7802
!HLA-A*02:03	2900
!HLA-A*02:06	2356
!HLA-A*02:16	697
!HLA-A*03:01	4553
!HLA-A*11:01	3463
!HLA-B*73:01	89
!HLA-C*04:01	487
!HLA-C*14:02	15
!HLA-C*15:02	72
HLA-A*02:01	9954
HLA-A*02:03	3485
HLA-A*02:06	3382

activity	unique_epi
HLA-A*02:16	223
HLA-A*03:01	3880
HLA-A*11:01	3824
HLA-B*73:01	43
HLA-C*04:01	35
HLA-C*14:02	228
HLA-C*15:02	109

Минимальный уровень дескрипторов - 6

Максимальный уровень дескрипторов - 16.

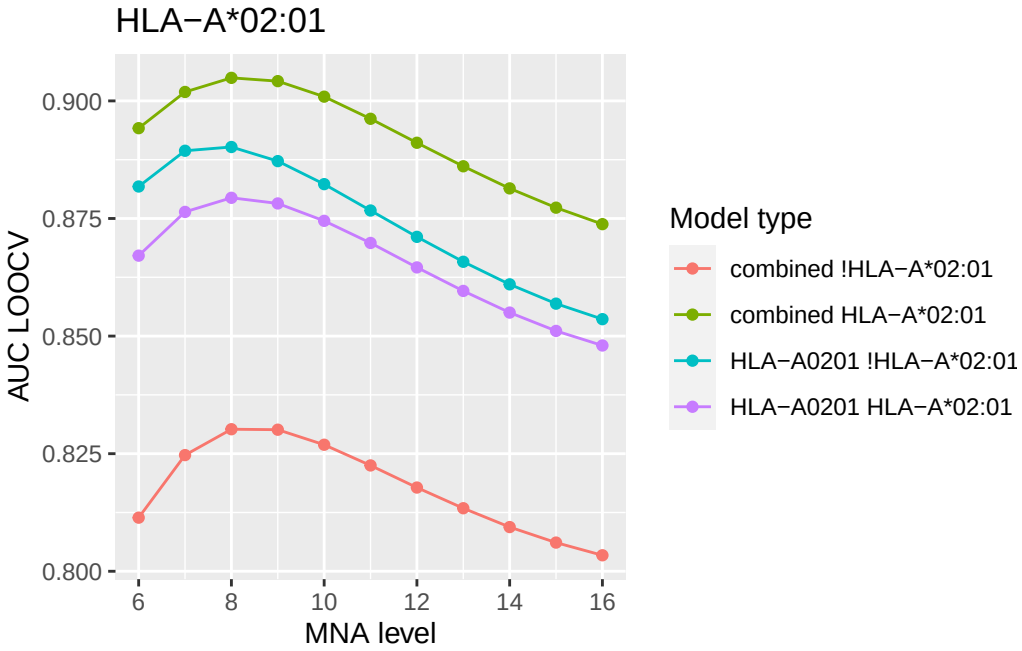
Из предыдущего опыта известно, что не стоит брать слишком маленький уровень дескрипторов, так как модели имеют низкую точность.

Результаты

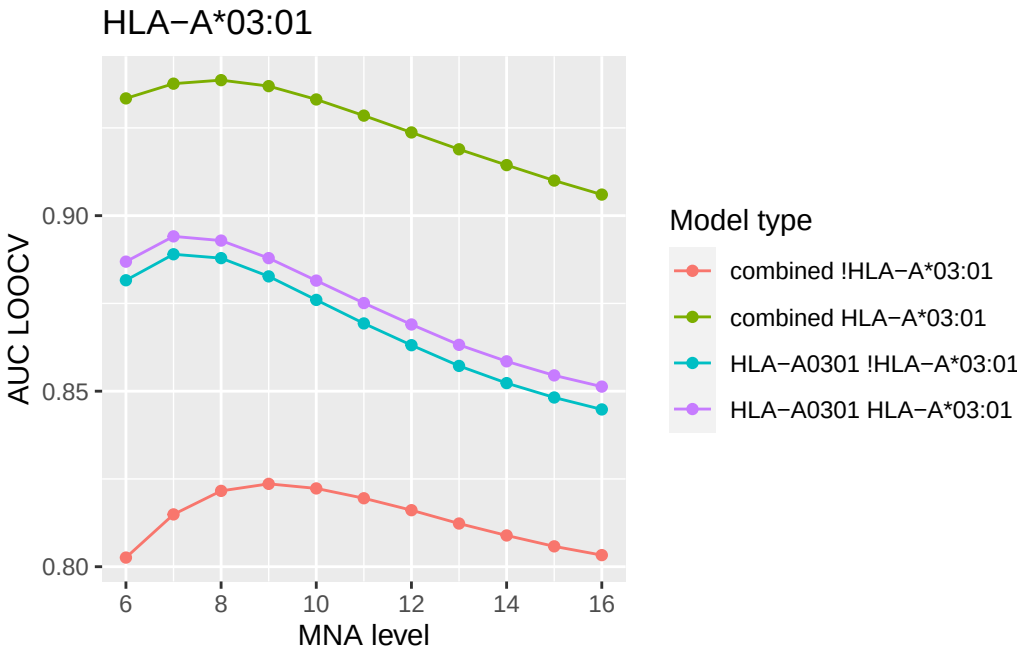
	model_name	descriptor_level	num_subst	iap	twentyCV	activity
1	combined	10	9954	0.9009	0.8990	HLA-A*02:01
2	combined	10	3821	0.9422	0.9414	HLA-A*11:01
3	combined	10	3878	0.9331	0.9322	HLA-A*03:01
4	combined	10	3485	0.8321	0.8269	HLA-A*02:03
5	combined	10	3382	0.8132	0.8092	HLA-A*02:06
6	combined	10	228	0.9412	0.9415	HLA-C*14:02

Высокопредставленные аллели

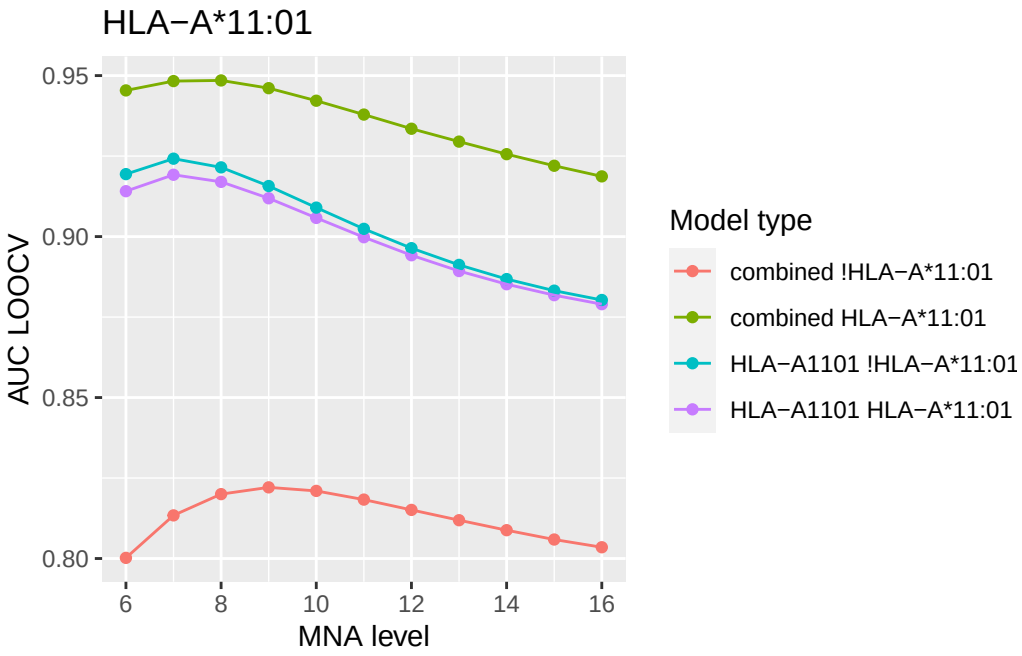
HLA-A*02:01



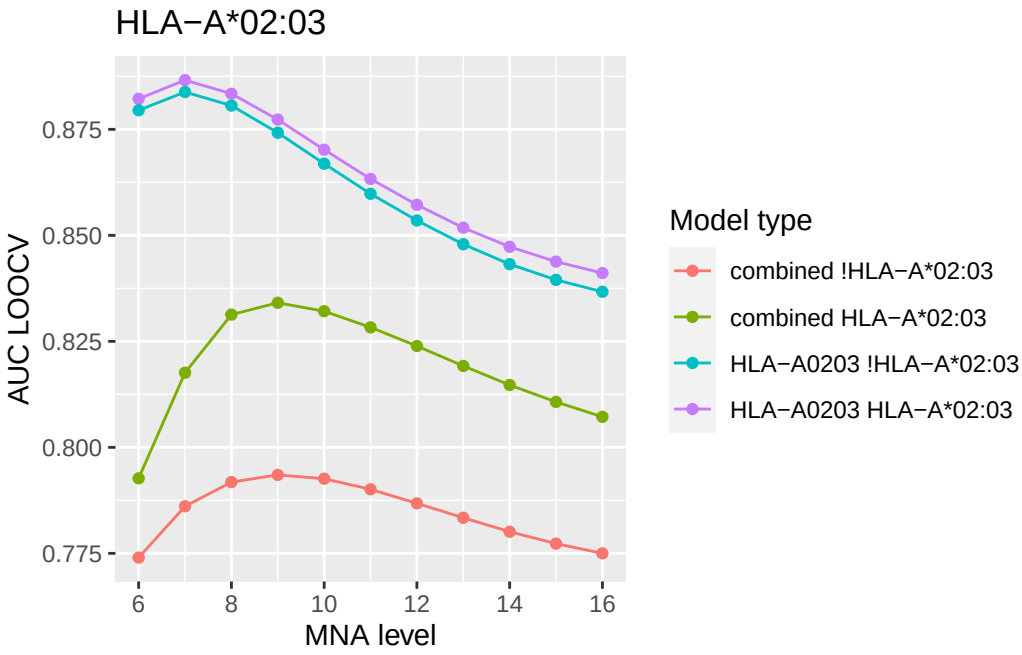
HLA-A*03:01



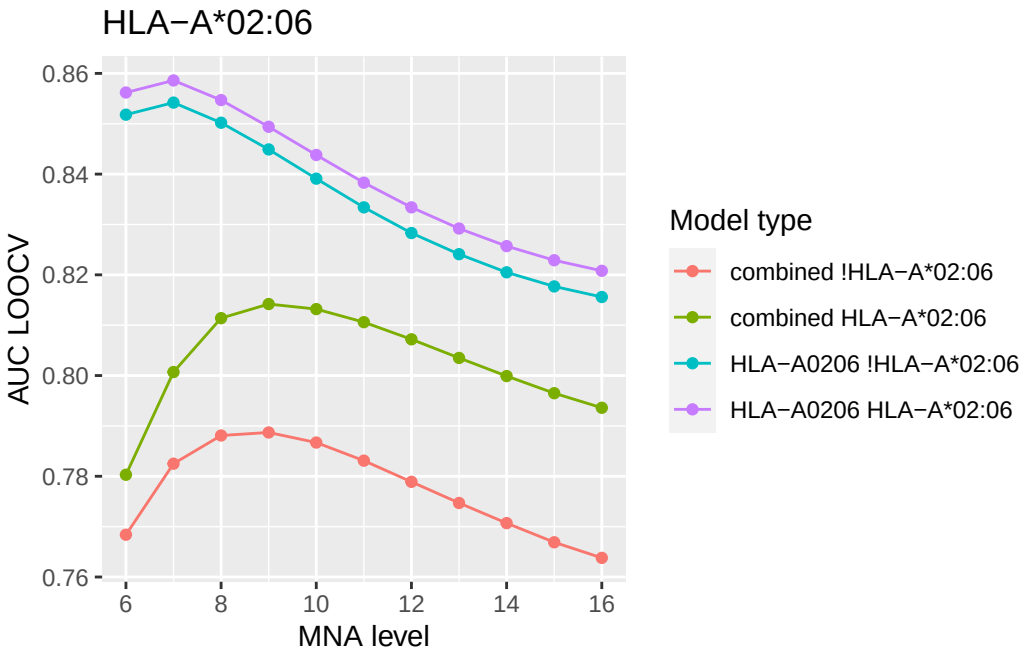
HLA-A*11:01



HLA-A*02:03

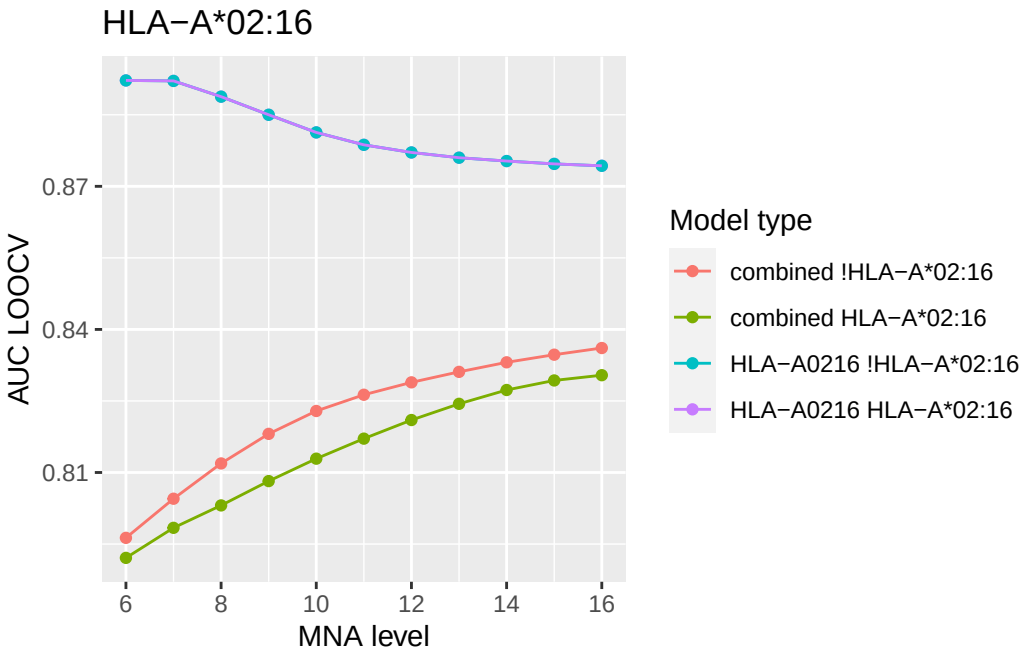


HLA-A*02:06

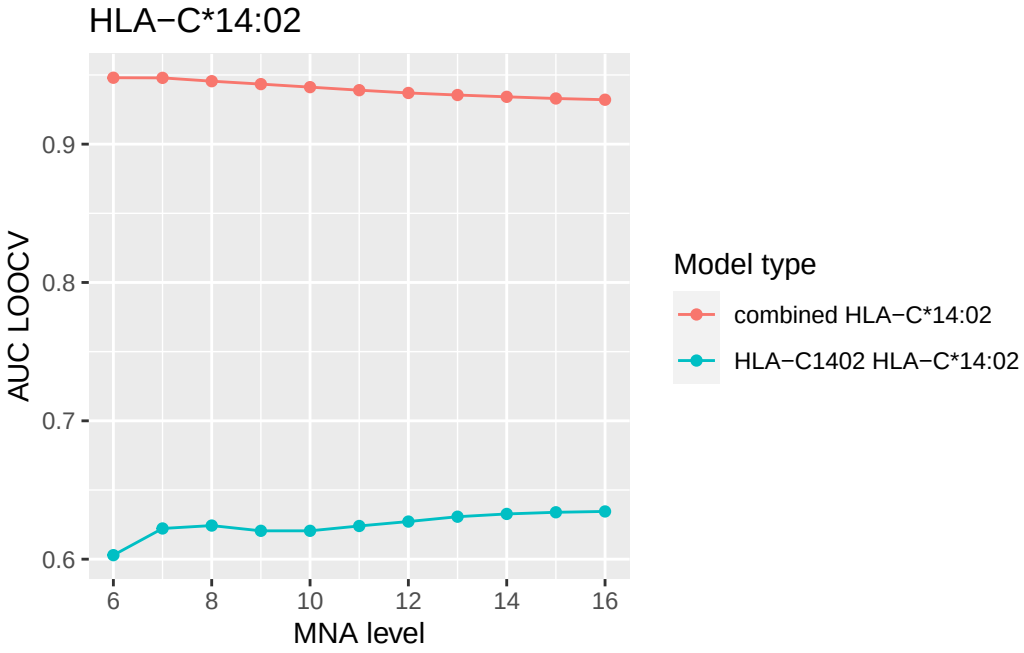


Низкопредставленные аллели

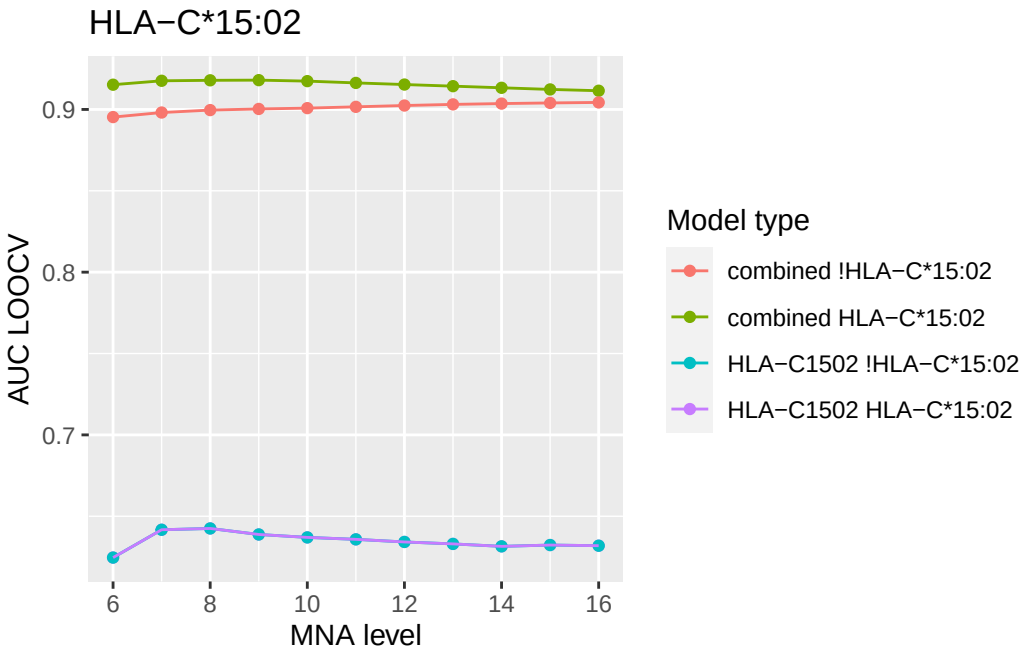
HLA-A*02:16



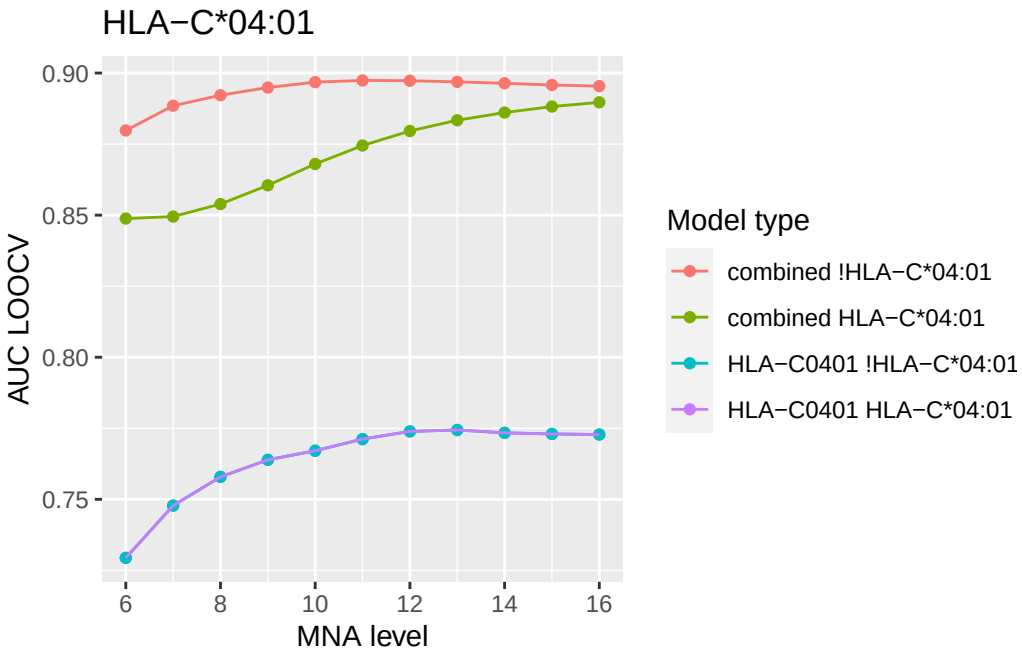
HLA-C*14:02



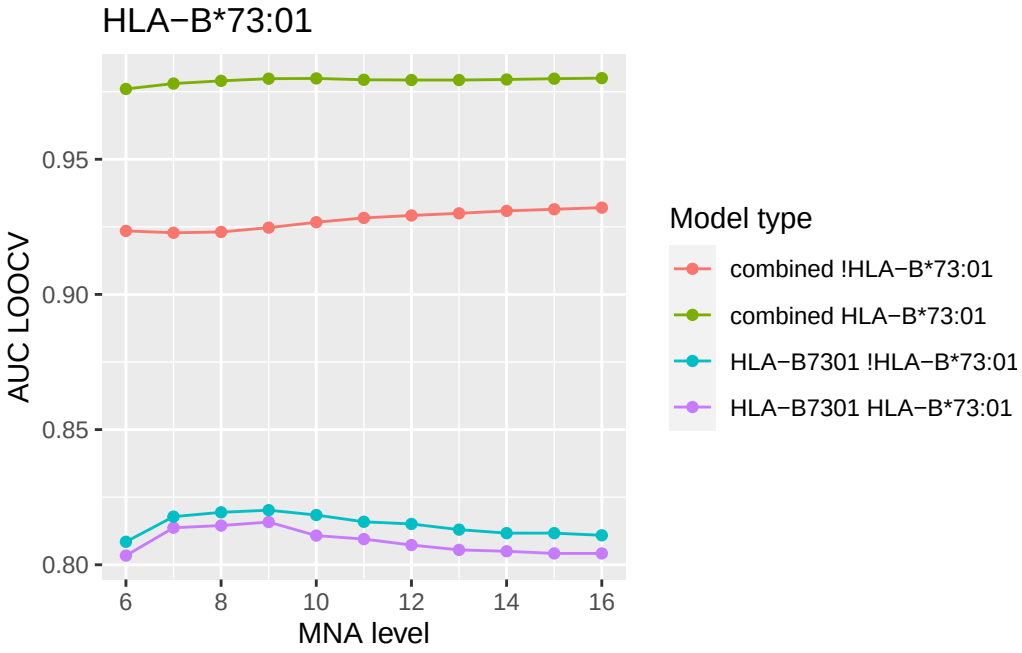
HLA-C*15:02



HLA-C*04:01



HLA-B*73:01



Обобщение

`summarise()` has grouped output by 'activity'. You can override using the `.groups` argument.

activity	model_name	max_auc	level
!HLA-A*02:01	HLA-A0201	0.8902	8
!HLA-A*02:01	combined	0.8302	8
!HLA-A*02:03	HLA-A0203	0.8838	7
!HLA-A*02:03	combined	0.7935	9
!HLA-A*02:06	HLA-A0206	0.8542	7
!HLA-A*02:06	combined	0.7887	9
!HLA-A*02:16	HLA-A0216	0.8922	6
!HLA-A*02:16	combined	0.8361	16
!HLA-A*03:01	HLA-A0301	0.8890	7
!HLA-A*03:01	combined	0.8236	9
!HLA-A*11:01	HLA-A1101	0.9242	7
!HLA-A*11:01	combined	0.8221	9
!HLA-B*73:01	HLA-B7301	0.8202	9
!HLA-B*73:01	combined	0.9321	16
!HLA-C*04:01	HLA-C0401	0.7744	13

activity	model_name	max_auc	level
!HLA-C*04:01	combined	0.8974	11
!HLA-C*15:02	HLA-C1502	0.6425	8
!HLA-C*15:02	combined	0.9043	16
HLA-A*02:01	HLA-A0201	0.8794	8
HLA-A*02:01	combined	0.9049	8
HLA-A*02:03	HLA-A0203	0.8866	7
HLA-A*02:03	combined	0.8341	9
HLA-A*02:06	HLA-A0206	0.8586	7
HLA-A*02:06	combined	0.8142	9
HLA-A*02:16	HLA-A0216	0.8922	6
HLA-A*02:16	combined	0.8304	16
HLA-A*03:01	HLA-A0301	0.8941	7
HLA-A*03:01	combined	0.9386	8
HLA-A*11:01	HLA-A1101	0.9192	7
HLA-A*11:01	combined	0.9485	8
HLA-B*73:01	HLA-B7301	0.8158	9
HLA-B*73:01	combined	0.9800	16
HLA-C*04:01	HLA-C0401	0.7744	13
HLA-C*04:01	combined	0.8897	16
HLA-C*14:02	HLA-C1402	0.6345	16
HLA-C*14:02	combined	0.9480	6
HLA-C*15:02	HLA-C1502	0.6425	8
HLA-C*15:02	combined	0.9180	9

type	mean_auc	model_level
allele	0.8298947	7
combined	0.8754947	9

`summarise()` has grouped output by 'type'. You can override using the `.groups` argument.

type	response	mean_auc	model_level
allele	neg	0.8411889	7
allele	pos	0.8197300	7

type	response	mean_auc	model_level
combined	neg	0.8475556	9
combined	pos	0.9006400	8

Выводы

1. Совмещенная обучающая выборка даёт в среднем большую точность, чем отдельные по аллелям в отношении положительного исхода, однако отрицательный исход модели предсказывают примерно одинаково точно. Оптимальная стратегия использовать совмещенную обучающую выборку
2. Оптимальный уровень дескрипторов лежит в промежутке от 7 до 9. Точность моделирования на 8 и на 9 уровне дескрипторов меняется, как правило, слабо.

Обсуждение

В исследование были намеренно не включенны данные, полученные в экспериментах, направленные на определение связывания эпитопа и MHC, а не элюированные от пациентов и определенные масс-спектрометрией. Создатели MHCflurry пишут, что такие данные зависят от предыдущих этапов процессинга, что вносит систематическую ошибку в модели. Поскольку масс-спектрометрические данные значительно увеличивают точность прогноза, они идут на компромисс и пробуют ограничивать количество таких данных, но насколько правильный такой подход?