Выбор источников данных для формирования обучающих выборок для моделей эпитоп-МНС.

Антон Смирнов

Mar 26, 2023

Оглавление

Введение **2**

Методы **3**

Результаты **4**

Выводы **6**

Список литературы **7**

Введение

Существует два типа экспериментов для получения данных о связывании эпитопа и MHC: определение связывания пептида и MHC и элюирование пептидов из MHC и определение их последовательности с помощью масс-спектрометрии. Второй метод высокопроизводителен, с его помощью получают большое количество данных, но это почти всегда положительные примеры. Данных, полученных первым методом в несколько раз меньше, но содержание положительных и отрицательных примеров примерно одинаковое. Цель этого эксперимента определить, какие данные: mhc_bind, mhc_elution, combined - стоит брать для построения конечной модели.

Методы

Источники данных: IEDB и MHCflurry.

Порог активности для масс-спектрометрических данных был взят <=100 nM в отличие от 5000 nM, который использовался ранее. Порог в 100 нМ взят из статьи(O’Donnell, Rubinsteyn, and Laserson 2020), порог 5000 нМ на основании изучения mhc_bind данных из IEDB.

Моделирование проводилось на 7-9 уровнях MNA дескрипторов в GUI версии MultiPASS. После обучения были отфильтрованы модели, для которых IAP < 0,75.

Результаты

Суммарная статистика

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

model_name	descriptor_level	num_activity	mean_iap	mean_20_fold
bind	7	179	0.8929380	0.8922128
bind	8	191	0.8840770	0.8820806
bind	9	201	0.8758602	0.8734338
combined	7	259	0.9071286	0.9062193
combined	8	261	0.9040720	0.9019705
combined	9	281	0.8915413	0.8896157
elution	7	151	0.9347589	0.9328285
elution	8	152	0.9299757	0.9291026
elution	9	152	0.9232605	0.9225908

Если оставить только положительные случаи

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

model_name	descriptor_level	num_activity	mean_iap	mean_20_fold
bind	7	97	0.9206278	0.9198753
bind	8	97	0.9165753	0.9144557
bind	9	94	0.9144000	0.9124149
combined	7	172	0.9323006	0.9315366
combined	8	172	0.9291297	0.9273169
combined	9	173	0.9222168	0.9198960
elution	7	149	0.9352128	0.9332725

model_name	descriptor_level	num_activity	mean_iap	mean_20_fold
elution	8	150	0.9304627	0.9295927
elution	9	150	0.9237220	0.9230507

Если оставить только отрицательные случаи

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

model_name	descriptor_level	num_activity	mean_iap	mean_20_fold
bind	7	82	0.8601829	0.8594902
bind	8	94	0.8505415	0.8486723
bind	9	107	0.8420028	0.8391888
combined	7	87	0.8573632	0.8561667
combined	8	89	0.8556461	0.8529865
combined	9	108	0.8424037	0.8411111
elution	7	2	0.9009500	0.8997500
elution	8	2	0.8934500	0.8923500
elution	9	2	0.8886500	0.8881000

Выводы

Может показаться, что модели основанные на чисто масс-спектрометрических данных имеют в среднем более высокую точность, чем остальные. Однако, это влияние недостаточно хорошей точности предсказания отрицательных случаев. Это может быть связано с недостаточным количеством данных относительно положительных активностей. Для положительных активностей точность не зависит от происхождения данных, отличается только спектр предсказаний: масс-спектрометрические данные расширяют почти вдвое. Для дальнейшего моделирования считаю разумным использовать комбинированные данные, моделирование проводить на 7 уровне MNA дескрипторов.

Список литературы

O'Donnell, Timothy J., Alex Rubinsteyn, and Uri Laserson. 2020. "MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing." *Cell Systems* 11 (1): 42–48.e7. https://doi.org/10.1016/j.cels.2020.06.010.