Прогнозирование связывания белков с лигандами на основе аминокислотных последовательностей и структур низкомолекулярных химических соединений

Карасев Д.А.,[1] Соболев Б.Н.,[1] Лагунин А.А,[1,2] Филимонов Д.А.,[1] Поройков В.В.[1]

[1] *Научно-исследовательский институт биомедицинской химии им. В.Н. Ореховича, Москва, Россия*
[2] *Российский Национальный Исследовательский Медицинский Университет им. Н.И. Пирогова*

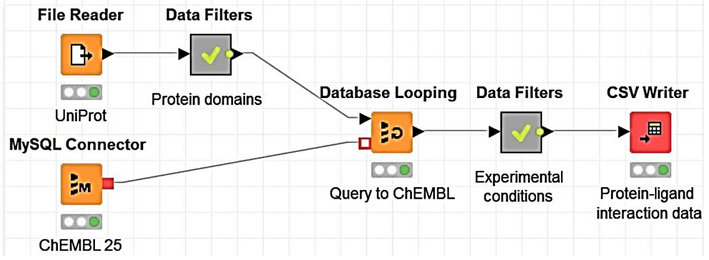
Введение

Компьютерное прогнозирование взаимодействий «белок-лиганд» - важный этап поиска фармакологических веществ. Полученные знания могут также служить для установления оптимальных направлений лекарственной терапии. В связи со значительной неполнотой экспериментальных данных по биологической активности фармакологических веществ при поиске белков-мишеней применяют методы машинного обучения. Цель работы заключается в разработке метода, который позволит эффективно прогнозировать взаимодействия для разнообразных групп белков и химических соединений.

Методы

1. Обучающие данные

Для извлечения информации о белках-мишенях и взаимодействующих с ними лигандами нами был разработан автоматизированный конвейер на платформе KNIME(https://www.knime.com).



Источником информации служила база данных ChEMBL. Отобраны результаты исследований in vitro, с указанием показателей аффинности IC50, Kd или Ki. Проведена фильтрация мутантных форм белков (таб. 1).

Таблица 1. Характеристики собранных данных

Группа белков		Activity cutoff (μM)	Ligand	Target
GPCR	IC50	1	15025	136
		10	17488	138
	Kd	1	557	65
		10	612	67
	Ki	1	32490	125
		10	35888	125
Ion Channel (LG*)	IC50	1	1840	17
		10	1980	22
	Kd	1	13	5
		10	15	5
	Ki	1	1018	21
		10	1165	22
Ion Channel (VG**)	IC50	1	4809	36
		10	8472	41
	Kd	1	7	6
		10	8	6
	Ki	1	976	4
		10	1485	4
Nuclear receptors	IC50	1	3832	29
		10	4827	29
	Kd	1	143	21
		10	168	22
	Ki	1	1134	23
		10	1238	24
Protein kinases	IC50	1	33078	277
		10	38795	292
	Kd	1	489	354
		10	574	356
	Ki	1	4077	115
		10	4410	123

2. Алгоритм

Использовался метод PASS, который был разработан для прогноза биологической активности лекарственно подобных органических соединений [1]. В данном методе используются оригинальные дескрипторы многоуровневых окрестностей атомов (MNA) (рис. 1), которые отражают локальные особенности взаимодействия лиганд-мишень. Метод основан на наивном байесовском классификаторе. Для каждой пары мишень-лиганд программа PASS рассчитывает оценки Pa и Pi, которые представляют собой вероятности взаимодействия и отсутствия взаимодействия между целевым белком и лигандом, соответственно.

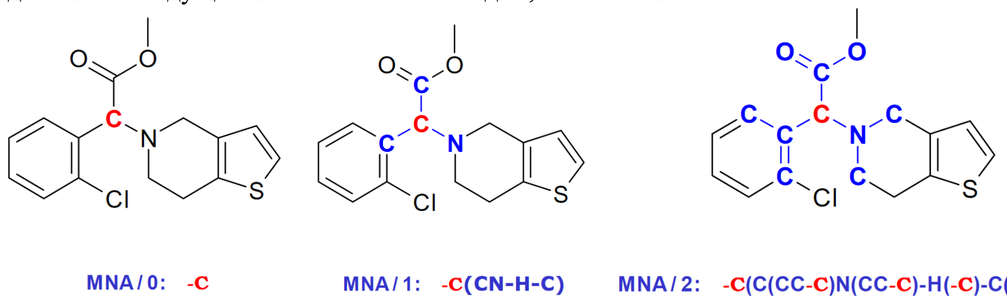


Рисунок 1. Пример генерации MNA-дескрипторов нулевого (MNA/0), первого (MNA/1) и второго (MNA/2) уровней.

Для оценки последовательностей использовалась программа SPrOS [2], в которой реализован алгоритм, основанный на оценках локального сходства аминокислотных последовательностей. Аминокислотная последовательность тестового белка (*Q*) сопоставляется с первичными структурами каждого из белков обучающей выборки (*K*) путем серии сдвигов.

```
AANRDPSQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVA  2
ANRDPSQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVAL  1
NRDPSQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALR  1
RDPSQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRA  0
DPSQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRAL  1
PSQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALF  2
SQFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFG  1
QFPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGR  1
FPDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRF  2
PDPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFP  0
DPHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPA  1
PHRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPAL  0
HRFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPALS  9
RFDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPALSL  0
FDVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPALSLG  3
DVTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPALSLGI  1
VTRDTRGHLSFGQGIHFCMGRPLAKLEGEVALRALFGRFPALSLGID  1


GTAINKPLSEKMMLFGMGKRRCIGEVLAKWEIFLFLAILLQQLEFSV  9
```

Последовательность *Q*

$$R_{ih} = \sum_{j=i}^{j+F-1} sim(q_j, k_{j+h})$$

K – последовательность обучающей выборки
Q – тестовая последовательность
i – позиция в тестовой последовательности *Q*
F – длинна сегментов сопоставления
sim(q,k) – подобие аминокислотных остатков в совмещенных позициях (идентичность)
h – величина сдвига между последовательностями *Q* и *K*

$$S_p = \max_{h,i} R_{ih},\ p - F < i \le p$$

Каждой позиции *p* присваивается наибольшее значение из оценок *Ri*

Оценка принадлежности белка к классам лигандной специфичности

$$t_p = \frac{\sum_{k=1}^{N} S_{pk} \times [a_k(C) - b_k(C)]}{\sum_{k=1}^{N} S_{pk} \times [a_k(C) + b_k(C)]}$$

C –лиганд обучающей выборки
tp – интегральная оценка для каждой позиции тестовой последовательности
Spk – оценка позиции *p* при сопоставлении с *k*-той последовательностью обучающей выборки
ak(C), *bk(C)* – коэффициенты принадлежности белка *k* к классу лигадной специфичности *C* и к его дополнению, соответственно

$$t = \sin\left[\frac{1}{m}\sum_{p=1}^{m} \arcsin(t_p)\right]$$

Усреднение оценок *tp* для всех позиций *m* тестовой последовательности *Q*

$$t_0 = \frac{\sum_{k=1}^{N} [a_k(C) - b_k(C)]}{\sum_{k=1}^{N} [a_k(C) + b_k(C)]}$$

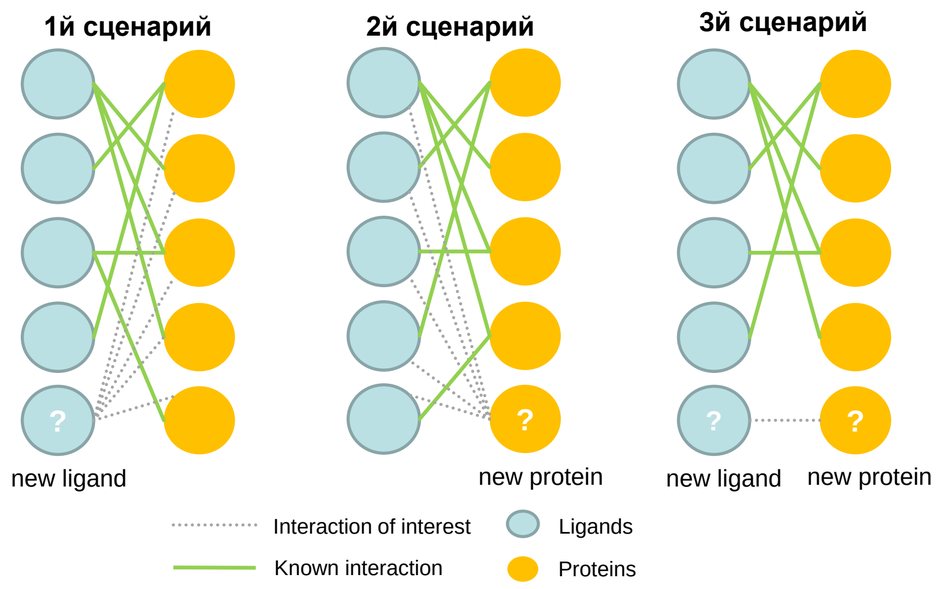
Для учета представительности класса в обучающей выборке рассчитывается *t0*

$$B(C) = \frac{t - t_0}{1 - tt_0}$$

Принадлежность белка *Q* к классу лигандной специфичности *C* оценивается с помощью *B*-статистики

Результаты

Тестирование программы осуществлялось на четырех наборах данных: рецепторы, связанные с G-белком (GPCR), протеинкиназы, ядерные рецепторы, ионные каналы (таб. 1). Тестирование осуществлялось в трех различных сценариях исследований (рис. 2).



При первом сценарии осуществляется прогноз связывания с белками, для которых уже известны лиганды. При втором – для нового белкам-мишени предсказывались лиганды, для которых уже известны какие-либо белки-мишени. Третий сценарий – для нового белка-мишени необходимо предсказать с новым лигандом.

Таблица 2. Результаты тестирования.

Группа белков	Показатель взаимодействия	Порог активности (μmol)	Первый сценарий (AUC)	Второй сценарий (AUC)	Третий сценарий (AUC)
GPCR	IC50	1	0,9864	0,968	0,904
		10	0,981	0,953	0,910
	Kd	1	0,9788	0,875	0,805
		10	0,9809	0,962	0,882
	Ki	1	0,9868	0,976	0,901
		10	0,9835	0,977	0,918
Протеинкиназы	IC50	1	0,9625	0,924	0,857
		10	0,954	0,902	0,838
	Kd	1	0,8079	0,790	0,655
		10	0,8026	0,797	0,650
	Ki	1	0,9797	0,956	0,893
		10	0,9806	0,937	0,856
Ion channel (ligand-gated)	IC50	10	0,9875	1,000	1,000
	Ki	1	0,9857	0,979	0,957
		10	0,9912	0,973	0,963
Ion channel (voltage-gated)	IC50	1	0,9857	0,985	0,932
		10	0,9677	0,898	0,839
Ядерные рецепторы	IC50	1	0,9843	0,973	0,852
		10	0,9811	0,988	0,943
	Kd	1	0,9686	1,000	0,961
		10	0,9715	1,000	0,972
	Ki	1	0,9926	0,988	0,976
		10	0,9955	0,987	0,989

Заключение

Таким образом, разработанный подход позволяет эффективно прогнозировать связывание белков с низкомолекулярными органическими соединениями. Высокая точность прогноза получена во всех сценариях прогноза, даже для пар белок-лиганд, когда ни для лиганда, ни для белка мишени ничего не известно. Наш метод применим для групп белков с различной степенью корреляции между филогенетическими отношениями и специфичностью к низкомолекулярным лигандам.

Работа выполнена при поддержке гранта РФФИ № 19-015-00374.

Ссылки:

1. Filimonov DA, Lagunin AA, Gloriozova TA, Rudik AV, Druzhilovskii DS, Pogodin PV, Poroikov VV, Prediction of the biological activity spectra of organic compounds using the PASS online web resource, *Chemistry of Heterocycl Compd* **50:** 444-457, 2014.
2. Karasev DA, Sobolev BN, Lagunin AA, Filimonov DA, Poroikov VV, Prediction of Protein–ligand Interaction Based on Sequence Similarity and Ligand Structural Features, *Int. J. Mol. Sci.* **2020**, *21*(21), 8152