XXVII Симпозиум «Биоинформатика и компьютерное конструирование лекарств»

Компьютерная оценка патогенности замен аминокислотных остатков на основе MNA дескрипторов пептидов отдельных белков и байесовского подхода

Задорожный А.Д[1], Смирнов А.С.[1],
Филимонов Д.А.[2], Лагунин А.А.[1,2]

[1] ФГАОУ ВО РНИМУ им. Н. И. Пирогова Минздрава России
[2] ФГБНУ «Научно-исследовательский институт биомедицинской химии имени В. Н. Ореховича

Актуальность

- Однонуклеотидные полиморфизмы (SNV, SNP) составляют более 80% от общего числа всех найденных генетических вариантов (База данных dbSNP содержит более 600 млн. записей);

- В белок-кодирующих областях генома несинонимичные варианты (nsSNV) меняют аминокислотные (АК) остатки. Они могут быть как безвредными, так и причиной нарушения функции белков (патогенными);

- Главными задачами медицинской генетики являются поиск и аннотация генетических вариантов. Современные технологии секвенирования генома позволяют обнаружить любые SNV, но определение их эффекта всё ещё сопровождает множество трудностей;

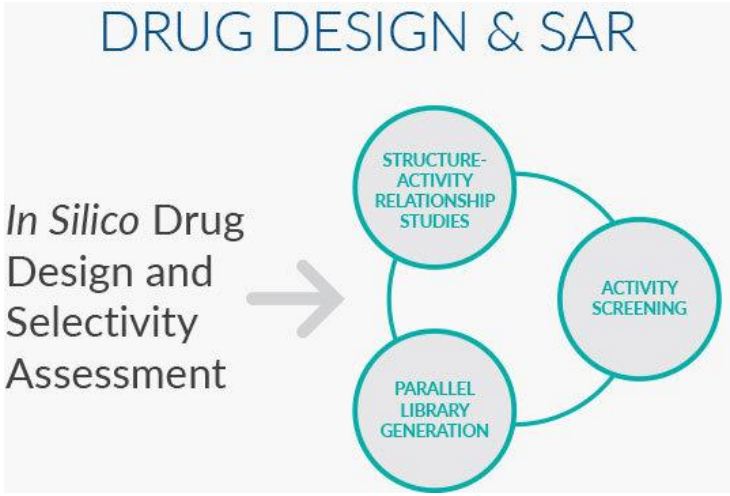
Актуальность

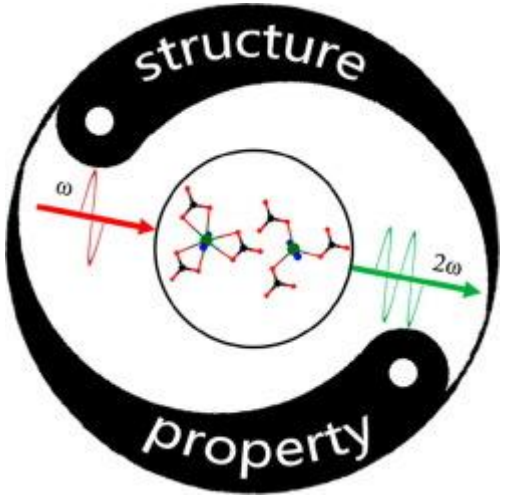
- В помощь исследователям и клиницистам создаются методы предсказания эффекта мутаций, в их основе лежат различные математические алгоритмы (нейронная сеть, скрытая марковская модель, метод опорных векторов и др.), свойства нуклеотидных/аминокислотных последовательностей (оптимальное выравнивание, заряд белка, частота в популяции и др.), системы оценивания (самостоятельные и комбинированные оценки);

- Однако, все существующие подходы используют для обучения большие гетерогенные данные, а также буквенное описание аминокислотных последовательностей;

- На данный момент нет общепринятого стандарта предсказания патогенности мутаций, каждый метод имеет преимущества и недостатки.

Цель исследования





В области компьютерного конструирования лекарственных средств себя хорошо зарекомендовали методы анализа связи структура-активность (SAR) и структура-свойство (SPR), и построение индивидуальных моделей, которые предсказывают взаимодействие вещества с конкретной мишенью на основе структурной формулы соединений.

Цель: Создать классификационную модель для прогноза патогенности аминокислотных замен на основе описания структурных формул фрагментов отдельных белков, в которых произошла мутация.

Материалы и методы

Известные аминокислотные замены взяты из баз данных:

ClinVar – содержит информацию о геномных вариациях и их связях со здоровьем человека

https://www.ncbi.nlm.nih.gov/clinvar/

Humsavar – перечень однонуклеотидных вариантов человека, собранный из литературных источников

https://www.uniprot.org/docs/humsavar

Uniprot – агрегирует надёжные, обширно описанные данные о различных белках

https://www.uniprot.org/

dbNSFP4.1a (academic version) – разработана для функционального прогнозирования и аннотации всех потенциальных несинонимичных однонуклеотидных вариантов (nsSNV) в геноме человека

http://database.liulab.science/dbNSFP

Название гена	*Связанная патология*
ATM	Атаксия телеангиоэктазия
ATP7B	Болезнь Вильсона
BRCA1	Рак груди и яичников
BRCA2	Рак груди и яичников
CFTR	Муковисцидоз
COL1A2	Несовершенный остеогенез
FBN1	Синдром Марфана
LDLR	Семейная гиперхолестеринемия
RYR1	Врожденные структурные миопатии
SCN5A	Синдром Бругада

Материалы и методы

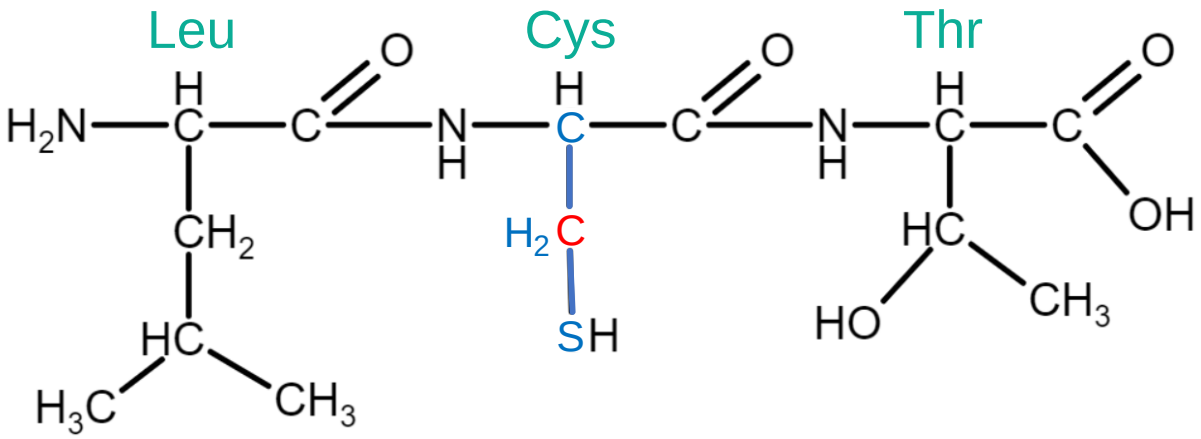
Программные средства – язык программирования Python 3.8 (библиотеки numpy, sklearn.metrics, Bio.SeqUtils, win32com.client); модифицированная версия Prediction of Activity Spectra for Substances (OnLineMultiPASS), оконная версия search_dbNSFP41a.jar

Метод машинного обучения – модифицированный наивный байесовский подход, используемый в программе PASS.

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Описание структуры – атомо-центрированные подструктурные дескрипторы многоуровневых атомных окрестностей (***M**ultilevel **N**eighborhoods of **A**toms* (**MNA**))

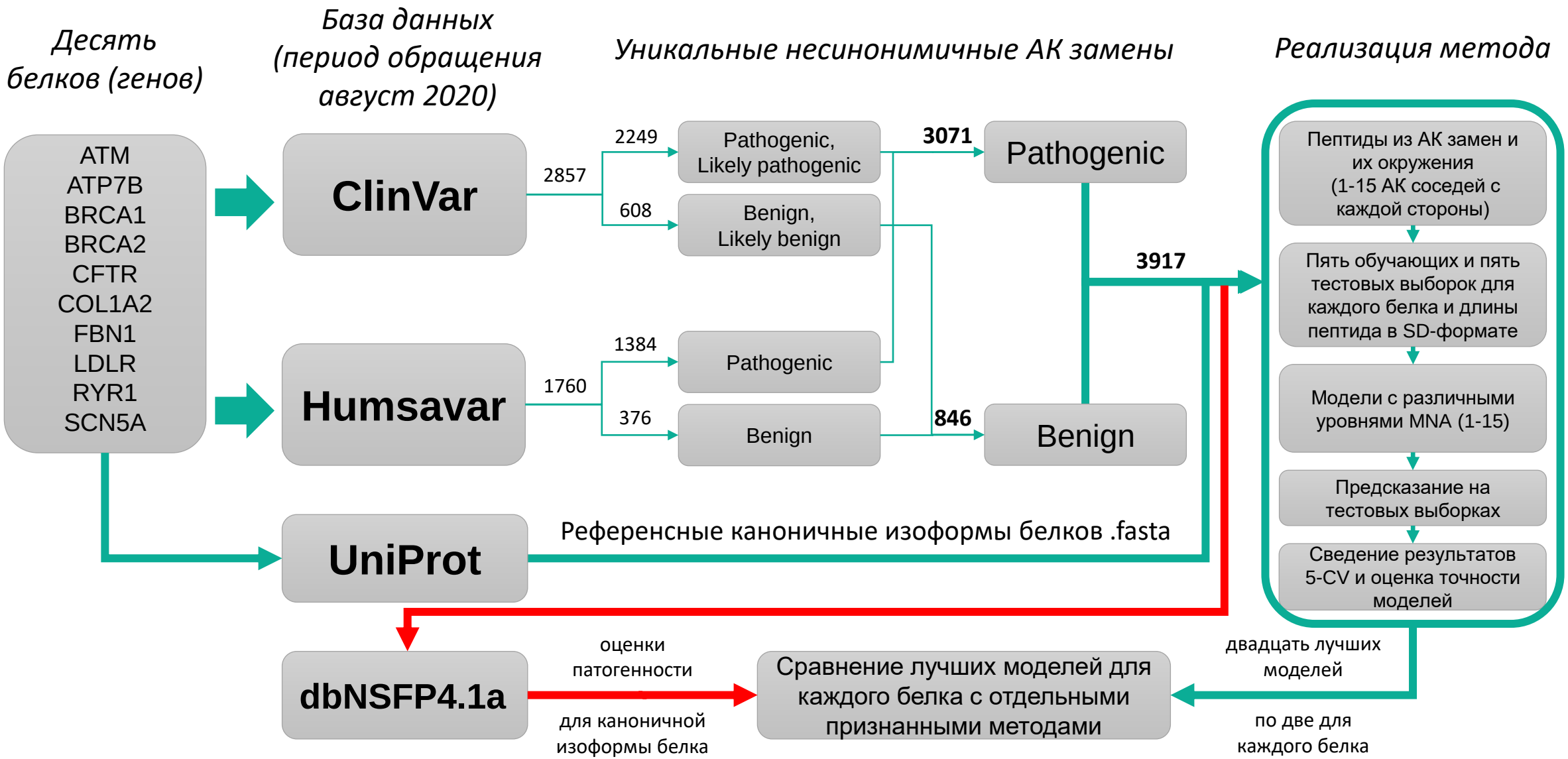
Пример для пептида : MNA/0: C
MNA/1: C(CSHH)



Filimonov D.A. et al. J. Chem. Inform. Computer Sci., 1999, 39, 666.

Karasev D.A., Savosina P.I., Application of molecular descriptors for recognition of phosphorylation sites in amino acid sequences, Biomed Khim., 2017 Oct;63(5):423-427

Результаты


Десять белков (генов)
ATM
ATP7B
BRCA1
BRCA2
CFTR
COL1A2
FBN1
LDLR
RYR1
SCN5A
База данных (период обращения август 2020)
ClinVar
Humsavar
UniProt
dbNSFP4.1a
Уникальные несинонимичные АК замены
2857
2249
Pathogenic, Likely pathogenic
608
Benign, Likely benign
1760
1384
Pathogenic
376
Benign
3071
Pathogenic
846
Benign
3917
Референсные канонические изоформы белков .fasta
оценки патогенности
для канонической изоформы белка
Сравнение лучших моделей для каждого белка с отдельными признанными методами
Реализация метода
Пептиды из АК замен и их окружения (1-15 АК соседей с каждой стороны)
Пять обучающих и пять тестовых выборок для каждого белка и длины пептида в SD-формате
Модели с различными уровнями MNA (1-15)
Предсказание на тестовых выборках
Сведение результатов 5-CV и оценка точности моделей
двадцать лучших моделей
по две для каждого белка

Результаты

Усредненные показатели точности предсказания эффекта АК замен в десяти клинически значимых белках:

Метод	Покрытие	$\overline{\text{ROC-AUC}}$	$\overline{\text{MCC}}$
MNA-based	99.5	0.808	0.447
PROVEAN	92.1	0.788	0.370
FATHMM	92.1	0.724	0.146
SIFT4G	89.3	0.816	0.432
PolyPhen2 HDIV	79.2	0.811	0.459
MutationAssessor	78.9	0.815	0.422

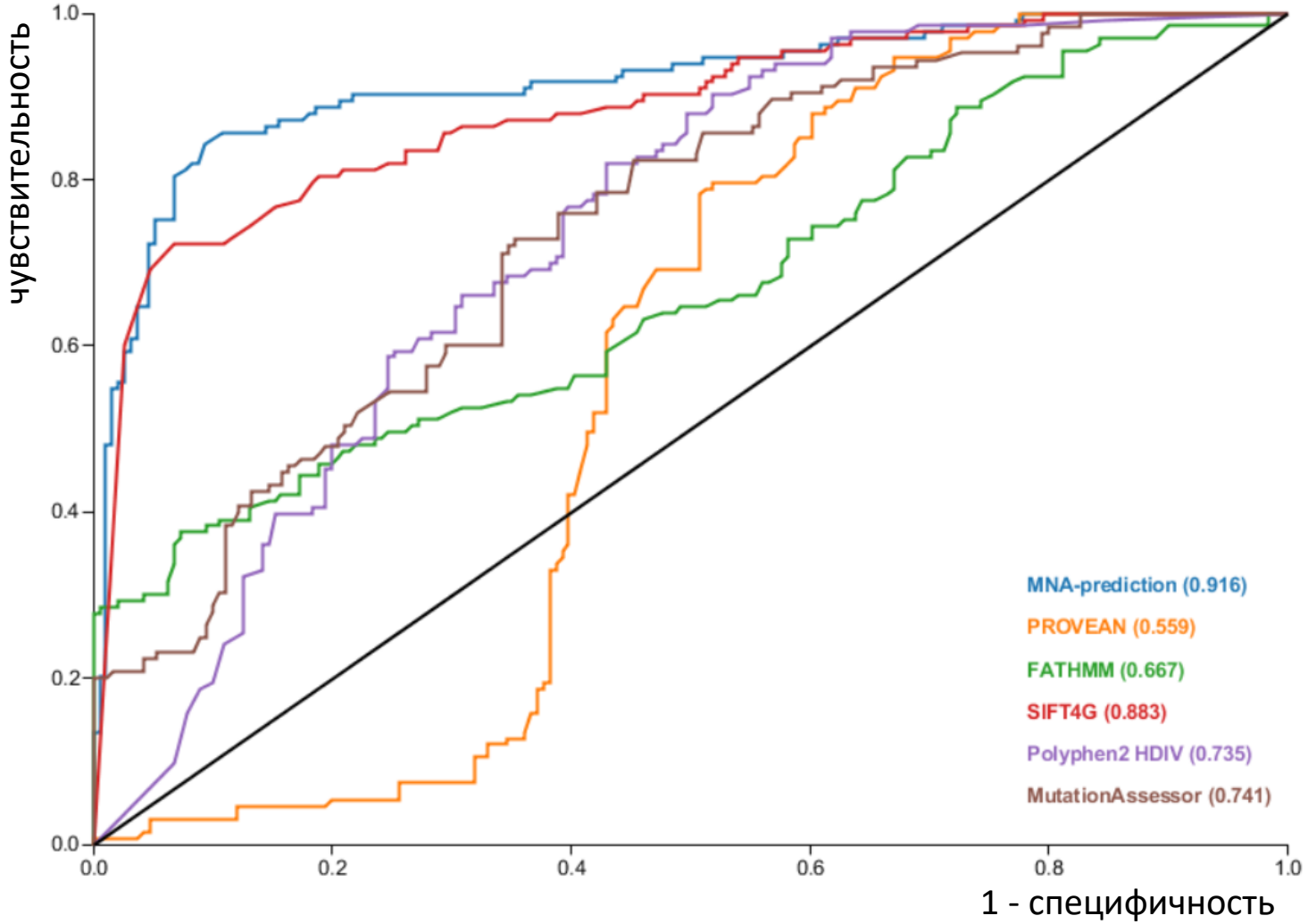
Покрытие(%) (3917 замен = 100%) относительно dbNSFP4.1a

Коэффициент корреляции Мэтьюса (MCC):

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

TP – истинно положительные
TN – истинно отрицательные
FP – ложно положительные
FN – ложно отрицательные

Сравнение ROC-кривых предсказаний эффекта АК замен в P38398 (BRCA1)



Заключение

Лучшие модели согласно коэффициенту корреляции Мэтьюса

Белок	Ген	MNA-based predictions				SIFT 4G		PROVEAN		PolyPhen 2 HDIV		MutationAssessor		FATHMM	
		Длина пептида	MNA	Покрытие(%)	MCC	Покрытие(%)	MCC	Покрытие(%)	MCC	Покрытие(%)	MCC	Покрытие(%)	MCC	Покрытие(%)	MCC
Serine-protein kinase ATM	ATM	21	14	100	0.217	97	0.547	96	**0.603**	96	0.500	96	0.453	96	0.215
Copper-transporting ATPase 2	ATP7B	23	12	100	0.474	100	0.451	100	**0.643**	100	0.607	100	0.345	100	0.160
Breast cancer type 1 susceptibility protein	BRCA1	31	13	100	**0.742**	99	0.490	99	-0.252	99	0.371	96	0.372	99	0.135
Breast cancer type 2 susceptibility protein	BRCA2	15	12	100	**0.420**	99	0.383	100	-0.018	0	–	0	–	100	0.368
Cystic fibrosis transmembrane conductance regulator	CFTR	15	10	100	0.225	100	0.231	100	**0.243**	100	0.218	100	0.194	100	0
Collagen alpha-2(I) chain	COL1A2	9	9	100	**0.891**	99	0.625	99	0.807	99	0.635	99	0.743	99	0.309
Fibrillin-1	FBN1	15	10	100	0.209	100	0.315	100	**0.364**	0	–	0	–	100	0.096
Low-density lipoprotein receptor	LDLR	9	14	100	0.317	99	0.493	99	**0.596**	99	0.583	99	0.490	99	0
Ryanodine receptor 1	RYR1	21	11	100	**0.560**	0	–	99	0.403	99	0.387	99	0.406	99	0.176
Sodium channel protein type 5 subunit alpha	SCN5A	15	9	100	**0.411**	99	0.355	28	0.314	99	0.372	99	0.371	28	0

1. Для каждого белка лучшая модель была получена при определенной длине пептида и уровне MNA дескрипторов.

2. Полученные значения MCC и AUC варьируют в широком диапазоне значений (ROC-AUC: 0.634–0.992, MCC: 0.209–0.891), что также наблюдается в результатах предсказаний сторонними методами.

3. Имеет место более высокая эффективность прогнозирования по отношению к FATHMM и отсутствие статистически значимых различий с остальными методами (U-критерий, p-value < 0,05).

Вывод: Результаты сравнения показателей точности прогноза говорят о конкурентоспособности подхода, основанного на MNA-дескрипторах и построении индивидуальных моделей, в предсказании патогенности аминокислотных замен и могут быть использованы в практических целях.

Спасибо за внимание!