**Колледж Научно-Технологического Университета Сириус**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**ДОКЛАД**

по дисциплине “Введение в специальность”

на тему “Применение нейронных сетей для создания вирусов”

Выполнил:  
Студент группы

К0709-23/1  
Бобров Даниил Анатольевич

Принял:

Старший преподаватель  
Тенигин Альберт Андреевич

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

IT-Колледж “Сириус”  
2023

1. **Вирусы**

Вирус. При этом слове люди думают о чем-то ужасном. Вирус – это болезнь, это боль, страдания, смерть. COVID19 только сильнее укрепил это в нашем подсознании. Но оказывается, есть вирусы, которые могут лечить. При чем лечат они от тех заболеваний, с которыми ни одно другое лекарство даже в теории справиться не сможет. Как же они это делают?

Давайте разберемся, из чего состоит вирус. Вирус – это неклеточная форма жизни. Если сильно упростить, то вирус состоит из последовательности ДНК и капсидов. Капсиды защищают ДНК, придают форму вирусу и отвечают за связывание вируса с клетками.

Смысл жизни вируса – найти клетку, связаться с ней и передать ей свое ДНК. Тогда клетка начнет производить копии этого вируса. Для связи с клетками вирус использует капсиды.

Но что, если заменить ДКН вируса? Впервые это было сделано в 1972 году. Тогда Пол Берг объединил ДНК обезьяньего вируса SV40 с ДНК лямбда-вируса.[1] Не было достоверно установлено, что этот вирус способен размножаться и/или заражать клетки. Но это был первый шаг в модификации вирусов. Сейчас ученые с легкостью справляются с этой задачей.

При чем, можно вставить в вирус любое ДНК. Даже ДНК человека. Тогда вирус будет стараться заразить клетки и распространить ДНК человека. Так можно использовать вирусы в генной терапии. Например, в 2005 году был создан первый вирус для лечения онкологии.[2]

Сейчас с помощью вирусов лечат сотни генетических заболеваний. Например, «Zolgensma» является самым дорогим лекарством в мире. Его стоимость установлена на уровне 2,125 млн долларов[3]. Это лекарство использует аденоассоциированный вирус 9 (AAV9) для доставки гена SMN1 в спинной мозг и лечения Спинальной мышечной атрофии (СМА)[4].

Но любой ли вирус нам подойдет? На самом деле нет. Дело в том, что жизнь вируса крайне опасна и недолговечна. Поэтому он может просто не успеть заразить какую-то клетку. Поэтому вирусы пытаются заражать все клетки, которые им встретятся. Но если, например, поместить ДНК печени куда-нибудь в легкие, то из этого явно не выйдет ничего хорошего.

А еще организм часто встречает вирусы, которые хотят его убить. Поэтому наша иммунная система зачастую убивает все вирусы и клетки, зараженные ими. Значит, нужен вирус, который не будет вызывать иммунную реакцию у организма.

Поэтому необходимо получить вирус, который будет атаковать лишь одну ткань организма и не будет вызывать иммунную реакцию. Также производство лекарств является бизнесом, поэтому целесообразно, чтобы вирус производился быстро, много и дешево.

Проанализировав все известные вирусы, ученые пришли к выводу, что самыми лучшими являются аденоассоциированные вирусы (AAV). Они удовлетворяют всем нашим требованиям: не вызывают иммунной реакции, точечно атакуют ткани нашего организма, они несут в себе достаточное количество ДНК и довольно просты в производстве.

Но даже AAV не идеальны. Например, упомянутая выше Золгенсма может повышать уровень ферментов печени и вызывать острое серьезное повреждение печени или острую печеночную недостаточность, что может привести к смерти [5]. Поэтому мы все еще продолжаем создавать новые виды AAV.

Но как понять, сколько тканей будет атаковать вирус?

Чтобы заразить клетку, вирусы используют капсиды. Они пытаются зацепиться за рецепторы на поверхности клетки. Но получается это не всегда. Давайте разберемся, когда вирусы справляются со своей миссией.

E(rN) = Ebonds + Eangles + Edihedrals + Ewdv + Ecoulomb

Ebonds

Eangles

Edihedrals

Ewdv

Ecoulomb

Эти 5 формул описывают взаимодействия атомов капсида вируса и рецептора клетки. Если мы посчитаем эти взаимодействия, то узнаем, сможет ли вирус заразить клетку.

С первыми тремя уравнениями у нас все хорошо. Просчитывать эти взаимодействия надо для каждого атома, поэтому их просчет не займет много времени. Проблема заключается в двух последних формулах. Их надо рассчитывать для каждой пары атомов. И это очень долго.

Капсид AAV состоит из капсомеров. Капсомер – это белок. Он состоит примерно из 730 аминокислот или 8000 атомов. И в одном капсиде 60 таких капсомеров. То есть чтобы посчитать одну конфигурацию AAV, надо просчитать примерно 100 миллионов взаимодействий. Это очень сложно, дорого и долго. Но возможно. Правда, таким способом возможно просчитать не более 5 аминокислот, а два вида AAV в среднем отличаются на 150 аминокислот. Необходимо перебрать все возможные наборы аминокислот, что невозможно. Это доказывает, что нам нужны новые методы создания вирусов.

Давайте поймем, как можно описывать капсиды. Каждый капсомер определяется последовательностью аминокислот, которая его задает. И она известна. Для каждого вируса возможно узнать эту последовательность, а затем узнать свойства, которыми обладает этот вирус. То есть можно построить некоторую функцию, которая будет для последовательности аминокислот возвращать набор свойств, которые она кодирует. Это довольно дорого, потому что вирус надо синтезировать, протестировать на клетках, а иногда и на животных, потом изучить свойства этих клеток. Но это возможно.

Только необходимо из свойств получить последовательность аминоклислот, которая будет кодировать идеальный вирус. И эта задача вызывает трудности. Подобные задачи, когда в одну сторону возможно что-то сделать, а в другую нет, очень часто встречаются в типичных задачах для генеративных нейронных сетей.

**2. Генеративно-состязательные нейронные сети**

Давайте ненадолго уйдем от вирусов, например, к фотографиям бобра.



Рис 1. Фотография Бобрика

Каждый человек легко может сказать, что изображено на этой картинке. Можно собрать большую обучающую выборку и сделать нейронную сеть, которая будет говорить, что это фотография бобра. Решить эту задачу очень просто. Чтобы это сделать, надо посадить много людей, дать им много фотографий, и они будут описывать эти фотографии. Но решать обратную задачу невероятно сложно. Но нейронные сети уже успешно решают и эту задачу. Давайте разберемся, как это возможно.

Самый первый успешный метод в 2014 году разработал Ян Гудфеллоу [6]. Это генеративно-состязательные нейронные сети. Давайте рассмотрим, как они работают.

Не известно, как разработать нейросеть, которая будет генерировать фотографии, потому что нет примеров того, что она должна выдавать на выход. Не существует примеров выходных данных, невозможно оценить ее точность.

Но можно для любой фотографии составить текстовое описание. Возьмем много фотографий бобра и много фотографий, на которых нет бобра. Сделаем нейронную сеть, которая будет отличать бобра от не бобра. Она будет отличать или дискриминировать одно от другого, поэтому такие нейросети называются дискриминаторами. Для нее надо просто максимизировать ответ на бобрах. То есть на фотографию бобра должен быть ответ 1, а на другие – 0.

Теперь надо создать вторую нейросеть – генератор. На вход будет подаваться описание, а на выходе будет получаться какое-то изображение. Теперь заморозим веса дискриминатора и будем учить эту нейросеть «обманывать» дискриминатор. То есть дискриминатор должен как можно чаще ошибаться и принимать фотографии, сгенерированные генератором, за фотографии бобра. И этот процесс будем повторять много раз, поочередно обучая то дискриминатор, то генератор.

В результате получим нейронную сеть, генерирующую фотографии бобра, имея только набор позитивных образцов. Прекрасно! И для картинок этот подход работает уже очень давно, но лишь в 2021 году этот подход решили применить для генерации белков [7].

Группа исследователей взяли 16 000 белков малатдегидрогеназ, ферментов длиной около 450 аминокислот, в два раза меньше, чем капсиды AAV. Эти ферменты обладают каталитической активностью, которую легко можно измерить и оценить качество работы нейросети. С помощью генеративно-состязательной нейросети сгенерировали 55 образцов и синтезировали их. 13 образцов реально обладали нужной каталитической активностью. При этом один из образцов на 106 аминокислот отличается от всех известных нам малатдегидрогеназ. Они не пытались задать какие-то конкретные свойства, то есть они не хотели сгенерировать конкретно фотографии бобра.

Давайте сделаем то же самое для AAV. Но есть проблема: если малатдегидрогеназ можно найти аж 16 000, то капсидов AAV лишь 170. Это очень грустно, но есть выход. Можно использовать не только капсиды аденоассоциированных вирусов. Для обучения нейросети подойдут и депендопарвовирусы, к которым относятся аденоассоциированные вирусы. Более того, можно использовать даже парвовирусы, предки депендопарвовирусов. И если собрать всю родню аденоассоциированных вирусов, то наберется около 20 000 экземпляров. Но ведь необходим конкретно AAV, а не любой парвовирус. Поэтому изменим дискриминатор, чтобы он не говорил, является ли последовательность AAV, а определял, каким типом парвовируса эта последовательность является. Тогда можно использовать для обучения сети все парвовирусы, ведь они очень похожи, но получать будем именно AAV.

Этот метод доказал, что можно генерирвоать аденоассоциированные вирусы с помощью нейронных сетей. Но ученые на этом не остановились и продолжили свои исследования.

**3. Диффузия**

Сейчас есть более эффективный метод генерации вирусов. Это диффузионные модели.

Включим генератор случайных чисел и создадим случайный шум. Дальше совместим нашу картинку с шумом. Получится зашумленное изображение. Этот процесс можно повторять итеративно несколько сотен раз. В результате получим настолько зашумленную фотографию, что она не будет отличаться от случайного шума. При чем, в датасете есть информация о том, как выглядело изображение на каждой итерации зашумления. 

Рис 2. Итеративное зашумление фотографии Бобрика

Теперь обучим нейронную сеть расшумлять это изображение. На вход сеть будет получать зашумленное изображение, номер шага и описание того, что должно быть изображено. Эта технология не нова. Например, в телефонах в ночном режиме уже давно фотография улучшается нейронными сетями. Идея диффузии в том, что используя одну и ту же нейронную сеть, нужно итеративно сотни раз расшумлять это изображение.

В результате нейронная сеть обучится генерировать из случайного шума фотографии бобра. Задача генерации картинок успешно решена.

Давайте вернемся к капсидам аденоассоциированных вирусов. Можно взять мишень, выбрать место, куда будет встраиваться вирус. Далее поместим туда случайную геометрию и с помощью диффузии будем расшумлять эту геометрию. В результате получим геометрию, которая будет связываться с клеткой в выбранном месте.

И 10 декабря 2022 года вышел прекраснейший инструмент, который решает эту задачу – RoseTTAFold Diffusion.

RoseTTAFold использует глубокое обучение для быстрого и точного прогнозирования белковых структур на основе только аминокислотных последовательностей. С помощью RoseTTAFold структуру белка можно вычислить всего за десять минут на одном компьютере.

RoseTTAFold одновременно отслеживает сразу три фактора: закономерности в последовательностях белков, взаимодействие аминокислот белка друг с другом, а также возможную трехмерную структуру белка. В этой архитектуре одно-, двух- и трехмерная информация передается взад и вперед, позволяя сети коллективно рассуждать о взаимосвязи между химическими частями белка и его структурой[8].

Диффузия – это расширение RoseTTAFold, которое уточняет предсказанные структуры. Для этого используется представление белковой основы в виде жесткого каркаса и нейросеть-трансформер SE(3)-equivariant, моделирующая движения атомов[9]. Диффузия также может генерировать многоатомные модели белков с расширенными функциями.

По результатам исследований, RoseTTAFold демонстрировал высокую точность в предсказании трехмерной структуры белков. Авторы заявляют, что RFdiffusion успешно решает 23 из 25 контрольных задач[10].

Теперь есть геометрия, но неизвестно, какая последовательность аминокислот свернется в эту геометрию. Эта задача уже давно успешно решается. Например, 4 июня 2022 года вышел инструмент ProteinMPNN[11].

ProteinMPNN – это метод, основанный на глубоком обучении, проектирования белковых последовательностей, который используется для поиска новых последовательностей аминокислот, которые будут складываться в определенную структуру белка. Этот метод использует нейронные сети передачи сообщений (MPNNs) для кодирования структуры и генерации последовательности. Он способен обрабатывать связанные позиции, когда аминокислота в одной позиции зависит от другой позиции.

Но оба алгоритма работают не очень хорошо. Их точность не идеальна. Поэтому исследователи предложили использовать модель AlphaFold2, чтобы проверить точность первых двух моделей.

Модель AlphaFold 2 построена на успешном сочетании глубокого обучения и нейронных сетей. Архитектура включает в себя генеративную и дискриминативную модели, использующие трансформерные нейронные сети. Эти модели анализируют последовательность аминокислот и преобразовывают ее в графовое представление, учитывая геометрию и взаимодействия в структуре белка. Далее модель делает предсказание координат атомов в трехмерном пространстве, формируя предложенную структуру белка.

AlphaFold 2 обучен на обширных данных белков, включая информацию о структурных базах данных и рентгеноструктурном анализе. Использование трансформеров позволяет модели улавливать долгосрочные взаимодействия в последовательностях аминокислот, что полезно для предсказания структуры белка. Многоуровневый подход модели позволяет анализировать белок на разных уровнях абстракции, учитывая множество взаимодействий и факторов, влияющих на его структуру.

Использование AlphaFold2 или RoseTTAFold Diffusion для оценки вероятности того, что спроектированная последовательность примет разработанную структуру мономера, и вероятность того, что эта структура связывается с мишенью, как было задумано, увеличивает вероятность успеха дизайна почти в 10 раз. Также обнаружено, что дизайн последовательности с использованием ProteinMPNN, а не Rosetta, повышает эффективность вычислений примерно в 8 раз[12][13].

Так исследователи научились генерировать новые виды вирусов, используя нейронные сети.

Хотя представленные здесь нейронные сети показывают реально большую эффективность и производительность, их точность все еще далека от идеала; показатели успеха среди мишеней меньше 1%[14]. Поэтому разработки нейронных сетей для генерации белков все еще продолжаются, а уже полученные результаты лишь увеличивают интерес к этим исследованиям.

**Список использованных источников**

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC389671/ – David A. Jackson, Robert H. Symons и Paul Berg. Биохимический метод внедрения новой генетической информации в ДНК вируса Simian Virus 40.

2. https://doi.org/10.1093/jnci/djj111 – Ken Garber, Китай одобрил первую в мире терапию онколитическим вирусом для лечения рака. Journal of the National Cancer Institute, том 98, выпуск 5, 1 марта 2006, страницы 298–300.

3. https://www.theguardian.com/science/2019/may/25/21m-novartis-gene-therapy-to-become-worlds-most-expensive-drug – Генная терапия Novartis стоимостью 2,1 миллиона долларов станет самым дорогим препаратом в мире.

4. https://www.zolgensma.com/how-zolgensma-works – Как работает Zolgensma.

5. https://www.zolgensma.com/zolgensma-studies – Результаты клинических исследований.

6. https://arxiv.org/pdf/1406.2661.pdf – Генеративно-состязательные нейронные сети

7. https://www.nature.com/articles/s42256-021-00310-5 – Expanding functional protein sequence spaces using generative adversarial networks

8. https://www.ipd.uw.edu/wp-content/uploads/2021/07/Baek\_etal\_Science2021\_RoseTTAFold.pdf – Accurate prediction of protein structures and interactions using a three-track neural network

9. https://www.biorxiv.org/content/biorxiv/early/2022/12/10/2022.12.09.519842/DC1/embed/media-1.pdf?download=false – Supplementary Methods for Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models

10. https://www.biorxiv.org/content/10.1101/2022.12.09.519842v1 – Широко применимый и точный дизайн белков за счет интеграции сетей прогнозирования структуры и диффузионно-генеративных моделей.

11. https://www.biorxiv.org/content/10.1101/2022.06.03.494563v1 – Надежный дизайн белковых последовательностей на основе глубокого обучения с использованием ProteinMPNN

12. https://www.biorxiv.org/content/10.1101/2022.06.15.495993v1 – Improving de novo Protein Binder Design with Deep Learning

13. https://www.biorxiv.org/content/10.1101/2022.06.03.494563v1.full – Robust deep learning based protein sequence design using ProteinMPNN

14. https://www.biorxiv.org/content/10.1101/2022.06.15.495993v1.full – Improving de novo Protein Binder Design with Deep Learning