**IT-Колледж «Сириус»**

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

**ДОКЛАД**

по дисциплине «Введение в специальность»

на тему «Биоинформатика и методы в биоинформатике»

Выполнил:  
Студент группы

1.9.7.1  
Дятлова Мария

Андреевна

Принял:

Старший преподаватель  
Тенигин Альберт Андреевич

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

# Содержание

[Содержание 2](#__RefHeading___1)

[Глоссарий 3](#__RefHeading___2)

[Введение 4](#__RefHeading___3)

[Актуальность направления 5](#__RefHeading___4)

[Языки программирования в биоинформатике 5](#__RefHeading___5)

[Алгоритмы Байера-Мура и Кнута-Морриса-Пратта 7](#__RefHeading___6)

[Роль глубокого обучения в математической биологии 8](#__RefHeading___7)

[Биоинформатика на примере одного исследования 9](#__RefHeading___8)

[Выводы и обсуждение 11](#__RefHeading___9)

[Список литературы 12](#__RefHeading___66)

[Приложение 13](#__RefHeading___11)

# Глоссарий

1. ЭКГ – электрокардиограмма
2. ЭЭГ – электроэнцефалограмма
3. МО – машинное обучение
4. ИИ – искусственный интеллект
5. DCM (dilated cardiomyopathy) – кардиомиопатия
6. Deep Learning – глубокое обучение
7. ARG – гены, связанные с аутофагией
8. LASSO - Оператор наименьшей абсолютной усадки и выбора
9. SVM-RFE - Устранение машинно-рекурсивной функции опорного вектора

# Введение

Биоинформатика (математическая биология) - это нестандартное направление, объединяющее на стыке как гуманитарные, так и технические специальности. Математическая биология включает в себя такие науки и предметы, как: общую биологию, генетику, химию, высшую математику, молекулярную биологию, большие данные и искусственный интеллект, а также - статистику.

Биоинформатика отличается тем, что при работе со статистикой и машинным обучением берутся данные природного или лабораторного происхождения.

В математической биологии при анализе последовательностей используются методы дискретной математики, статистики и моделирования на основе математических способов.

В последнее время математическая биология вышла за пределы анализа только биологических последовательностей и стала важным инструментом в исследовании других видов био-данных – научных текстов, изображений (фотографий, микроскопических и томографических снимков), физиологических и биохимических данных, данных ЭЭГ и ЭКГ и многих других.  
В последние годы в биоинформатике все более активно используются методы машинного обучения и анализа больших данных. В реферате представлены и описаны некоторые алгоритмы, а также приведено исследование с методами математической биологии и машинного обучения, которые помогли выявить 10 генно-целевых препаратов и осложнений для DCM.

# Актуальность направления

Актуальность направления заключается в том, что область наук о жизни накапливает с каждым годом все больше и больше массивов данных, а направление, зачастую, нуждается в грамотном владении компьютерными методами и специальными алгоритмами.

Успех Финкбейнера подчеркивает, как глубокое обучение, одна из самых перспективных отраслей искусственного интеллекта (ИИ), вторгается в биологию. В биологии алгоритмы глубокого обучения погружаются в данные так, как люди не могут, обнаруживая функции, которые в противном случае было бы невозможно поймать. Исследователи используют алгоритмы для классификации клеточных изображений, установления геномных связей, продвижения открытия лекарств и даже поиска ссылок по различным типам данных, от геномики и визуализации до электронных медицинских записей.

# Языки программирования в биоинформатике

Наиболее часто используемыми в этой области являются языки программирования Python и R. Языки Python и R предназначены для работы с данными, и у обоих языков есть свои достоинства и недостатки. Выбор того или иного языка зависит от конкретной ситуации, а также от того, какие еще распространенные инструменты требуются для решения задачи.

**Python и R** – популярные языки программирования для работы со статистикой. В то время, как язык R разрабатывался с целью применения именно на потребности статистиков, Python более понятен синтаксисом. В настоящее время язык R является одним из ведущих статистических инструментов в мире. Он активно применяется в генетике, молекулярной биологии и биоинформатике, науках об окружающей среде (экология, метеорология) и сельскохозяйственных дисциплинах. Также R используется в обработке медицинских данных.

**Язык R** – интерпретируемый язык программирования, который работает в основном с помощью командного интрепретатора. Используется для обработки статистических данных и для работы с графикой. Но, помимо этого, выполняет такой функционал свободной программной средой вычислений с открытым исходным кодом. За счет установки дополнительных пакетов обеспечивает максимальную комфортность в использовании.

**Язык Python** – один из высокоуровневых языков программирования, является объектно-ориентированным, а сам синтаксис ядра минималистичен. Известен как интерпретируемый язык.

# Алгоритмы Байера-Мура и Кнута-Морриса-Пратта

Данные **два алгоритма** будут рассматриваться как те, которые даются в ВУЗ-ах для вводного изучения курса биоинформатики, без которых, собственно, сложно понимать тактику поиска биологических данных и работу с ними.

Большинство лабораторных работ подразумевает себя работу с алгоритмами, которые бы в процессе машинного обучения предсказывали следующую последовательность геномного кода. В тексте будут представлены приложения к простой реализации данных алгоритмов, без машинного обучения, на языке программирования Python.

**Алгоритм Бойера-Мура** — эффективный алгоритм общего вида, предназначенный для поиска подстроки в строке. Разработан Робертом Бойером и Джеем Муром в 1977 году. В биоинформатике используется для поиска определенных закономерностей в геномном коде и восстановление поврежденных строк предназначения геномного кода.

**Алгоритм Кнута - Морриса - Пратта** — алгоритм, осуществляющий поиск подстроки в строке. Время работы алгоритма линейно зависит от объёма входных данных. Алгоритм был разработан Д. Кнутом и В. Праттом и, независимо от них, Д. Моррисом, а результаты своей работы они опубликовали совместно в 1977 году.

# Роль глубокого обучения в математической биологии

Алгоритмы глубокого обучения полагаются на нейронные сети, вычислительную модель, впервые предложенную в 1940-х годах, в которой слои нейроноподобных узлов имитируют то, как человеческий мозг анализирует информацию.

Еще около пяти лет назад алгоритмы машинного обучения, основанные на нейронных сетях, полагались на исследователей для обработки необработанной информации в более значимую форму. Но взрыв в размере наборов данных - из таких источников, как снимки смартфонов или крупномасштабное геномное секвенирование - и алгоритмические инновации теперь позволили людям сделать шаг. Этот прогресс в машинном обучении - "глубокая" часть - заставляет компьютеры, а не их программистов-людей, находить значимые отношения, встроенные в пиксели и базы. И поскольку слои в нейронной сети фильтруют и сортируют информацию, они также взаимодействуют друг с другом, позволяя каждому слою уточнить вывод из предыдущего.

На Рис.1 один из примеров, как обучается нейросеть в процессе машинного обучения на основе биологических данных.

# Биоинформатика на примере одного исследования

Машинное обучение и биоинформатика для выявления 8 биомаркеров, связанных с аутофагией, и построения регуляторных сетей генов при расширенной кардиомиопатии. По собственным соображениям - исследовательская биоинформатика за отсутствием в некоторых исследованиях программистов почти всегда прибегает к случаям использования автоматизированных систем. В этом исследовании была предпринята попытка идентифицировать гены, связанные с аутофагией (ARG), с диагностическими биомаркерами DCM с использованием подходов машинного обучения и биоинформатики.

Хорошо известно, что DCM нарушает расширение желудочков и систолическую диастолу, что приводит к аритмиям и сердечной недостаточности в тяжелых случаях. К сожалению, при низкой распространенности EMB большинство пациентов с кардиомиопатией на ранней стадии не лечатся эффективно. Золотым стандартом для миокардита и DCM часто является плохой прогноз в случаях сопутствующей аритмии и сердечной недостаточности. Поэтому ранняя диагностика, точная оценка и терапевтическое лечение пациентов с DCM кажутся решающими. Следовательно, исследователи все чаще ищут диагностические маркеры DCM. Между тем, молекулярный патогенез DCM, вирусных инфекций и других факторов прогрессирования и прогнозирования заболевания все еще не полностью изучен. Дифференциальный анализ данных микрочипов целых генов DCM из базы данных Gene Expression Omnibus (GEO) проводился с использованием платформы NetworkAnalyst 3.0. Сопоставляется дифференциально экспрессированные гены были получены в наборе данных GSE4172 путем слияния ARG из библиотек генов аутофагии HADb и HAMdb для получения связанных с аутофагией дифференциально экспрессированных ген. Корреляционный анализ AR-DEG и их визуализация были выполнены с использованием языка R.

В настоящем исследовании ученые определили восемь молекулярных маркеров DCM, которые были получены с помощью двух алгоритмов машинного обучения. Создана регуляторная сеть факторов транскрипции. Наконец, было выявлено 10 генно-целевых препаратов и осложнений для DCM.

Для исследования были произведены скрининг диагностических биомаркеров для DCM с помощью алгоритмов машинного обучения и скрининг оптимальных генов с помощью регрессионной модели LASSO. С помощью МО, был построен график лучшего гена, выбранного алгоритмом SVM-RFE. (Рис.2)

# Выводы и обсуждение

В ходе данной работы было проведено ознакомление с базовыми понятиями биоинформатики, описаны некоторые алгоритмы, а также приведено исследование с методами математической биологии и машинного обучения, которые помогли выявить 10 генно-целевых препаратов и осложнений для DCM и его раннего диагностирования с помощью модели машинного обучения.

Помимо этого, было приведены популярные языки программирования, которые используются в сфере математической биологии.

В обсуждении стоит учесть то, что математическая биология является перспективным направлением в научно-технической сфере. В последующем, люди связанные с данным направлением поспособствуют развитию механизации биологических исследований, что является большим прогрессом в науке.

# **Список литературы**

1. 25 профессий, к которым готовит НИУ ВШЭ // НИУ ВШЭ. - 2022
2. Fengjun Zhang, Mingyue Xia, Jiarong Jiang, Shuai Wang, Qiong Zhao, Cheng Yu, Jinzhen Yu, Dexian Xian, Xiao Li, Lin Zhang, Yuan Liu & Min Peng Machine learning and bioinformatics to identify 8 autophagy-related biomarkers and construct gene regulatory networks in dilated cardiomyopathy // Nature. - 2022
3. Г. П. ВОЛИНЕЦЬ, В. Г. БДЖОЛА, С. М. ЯРМОЛЮК МЕТОДИ СТРУКТУРНОЇ БІОІНФОРМАТИКИ // ОГЛЯДИ. - 2010
4. Джассал, Б. и др. База знаний реакционного пути. Нуклеиновые кислоты Res. 48, D498-D503. - 2020.
5. Макнелли, Э. М. & Mestroni, L. Дилированная кардиомиопатия: генетические детерминанты и механизмы. Циркуляр Рез. 121, 731–748. - 2017.
6. О.Е. Баксанский БИОИНЖЕНЕРИЯ И БИОИНФОРМАТИКА: КОНВЕРГЕНТНЫЕ ТЕХНОЛОГИИ // Сеченовский вестник. - 2015. - №1(19). - С. 50-55.
7. Фридман, Дж., Хасти, Т. Тибширани Пути регуляризации обобщенных линейных моделей с помощью координатного спуска. // Дж. Стат. Софтв. - 2010. - С. 33.

# Приложение

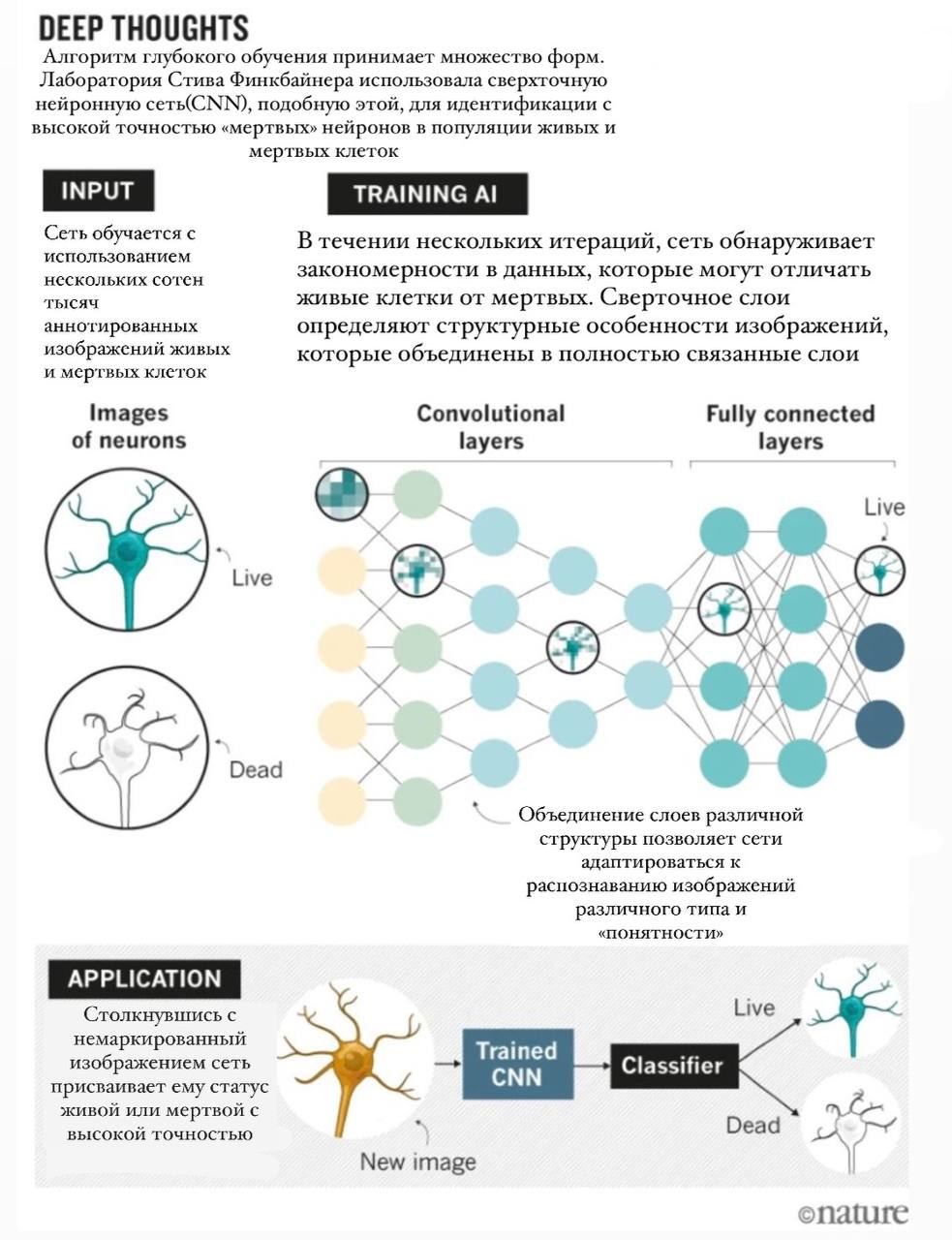


Рис.1 – Обучение нейросети компьютерному зрению на примере «мертвых» и «живых» клеток.

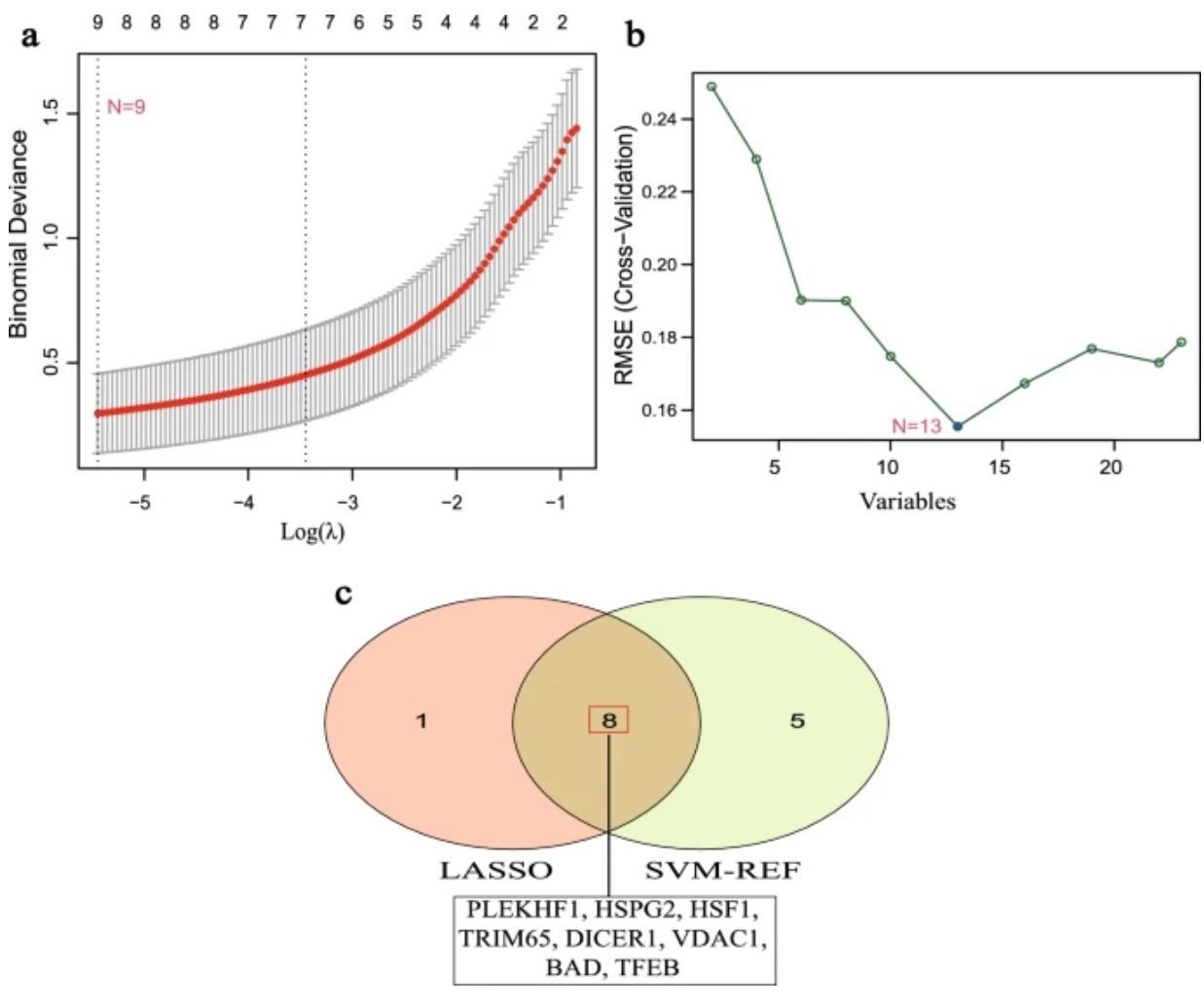


Рис.2 - График лучшего гена, выбранного алгоритмом SVM-RFE.