

**федеральное государственное автономное образовательное учреждение
высшего образования
ПЕРВЫЙ МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ МЕДИЦИНСКИЙ
УНИВЕРСИТЕТ имени И.М. СЕЧЕНОВА
Министерства здравоохранения Российской Федерации
(Сеченовский Университет)**

**Институт фармации им. А.П. Нелюбина
Кафедра биотехнологии**

**Мусина Лилия Ильшатовна
студентка 5 курса 709-02 группы**

**Потенциальные источники эндогенного пептида НАЕЕ
в организме человека**

ДИПЛОМНАЯ РАБОТА
по направлению подготовки (специальности)
06.05.01 Биотехнология и биоинформатика

**МОСКВА
2026**

Работа выполнена на базе Первого МГМУ им. И.М. Сеченова, кафедра информационных технологий и обработки медицинских данных.

И.о. зав. кафедрой биотехнологии,
доктор биологических наук

Костюшев Д.С.

Научный руководитель дипломной работы:

доцент кафедры информационных технологий
и обработки медицинских данных института
цифрового биодизайна и моделирования
живых систем Сеченовского университета,
кандидат физико-математических наук

Анашкина А.А.

С дипломной работой можно ознакомиться на кафедре биотехнологии Института фармации им. А.П. Нелюбина Сеченовского университета по адресу: пр-т Вернадского, д. 96, корп. 1.

Оглавление

ВВЕДЕНИЕ.....	5
1. ОБЗОР ЛИТЕРАТУРЫ.....	7
1.1. Болезнь Альцгеймера.....	7
1.2. Молекулярные основы болезни Альцгеймера.....	8
1.3. Генезис β -амилоидного пептида.....	10
1.4. История разработки пептида НАЕЕ.....	12
1.5. Открытие эндогенности НАЕЕ.....	13
1.6. Гипотезы источника НАЕЕ в организме человека.....	13
1.7. Постановка задачи.....	14
2. МАТЕРИАЛЫ И МЕТОДЫ.....	18
2.1. Материалы.....	18
2.2. Методы.....	18
2.2.1. Фильтрация и аннотирование найденных последовательностей.....	18
2.2.2. Идентификация открытых рамок считывания в межгенных и 5'UTR регионах.....	20
2.2.3. Статистический анализ.....	20
2.2.4. Биоинформатическое программное обеспечение и базы данных.....	20
2.2.5. Валидация результатов с помощью BLASTN.....	21
3. РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ.....	22
3.1. Поиск вырожденного мотива в геноме человека.....	22
3.1.1. Аннотирование и фильтрация найденных последовательностей.....	24
3.1.2. Контрольная валидация результатов с использованием BLASTN.....	33
3.1.3. Анализ расположения мотивов относительно сплайсируемых участков.....	35
3.2. Анализ дополнительных геномных областей, содержащих мотив пептида НАЕЕ.....	38
3.2.1. Анализ межгенных регионов.....	39
3.2.2. Анализ 5'UTR регионов и идентификация uORF.....	42
3.2.2.1. Контрольная валидация результатов с использованием BLASTN.....	46
ВЫВОДЫ.....	47
СПИСОК ЛИТЕРАТУРЫ.....	48
СПИСОК СОКРАЩЕНИЙ.....	53
Приложение А.....	54
Результаты BLASTN-валидации.....	54
Приложение Б.....	57

Анализ локализации мотива НАЕЕ в неcodирующих регионах.....	57
---	----

ВВЕДЕНИЕ

Актуальность исследования. Болезнь Альцгеймера является наиболее распространенной формой нейродегенеративной патологии, которая характеризуется постепенным снижением когнитивных навыков, таких как память, внимание и мышление [1, 4].

Несмотря на значительное накопление знаний в области нейродегенеративных заболеваний, все же остаются вопросы о молекулярных механизмах патогенеза, которые являются предметом активных исследований. В центре внимания патогенеза заболевания лежит явление образования и агрегации β -амилоидного пептида (A β), формирующего внеклеточные амилоидные бляшки [6,8,11]. Благодаря развитию пептидомики и масс-спектрометрического анализа микробелков исследователи обратили внимание на короткие пептиды, которые могут быть взаимосвязаны с агрегацией A β и оказывать нейропротекторное действие. Одним из таких соединений является N-ацилированный и C-амидированный тетрапептид HАЕЕ (His-Ala-Glu-Glu), который первоначально разрабатывался как синтетический пептид с антиагрегатной активностью [32,36,31]. Позднее появились данные, указывающие на эндогенность HАЕЕ в организме человека [37,38]. В связи с этим актуальной задачей является поиск возможных источников и путей образования пептида HАЕЕ в организме человека.

Целью данной работы является изучение происхождения эндогенного пептида HАЕЕ в организме человека. Для выполнения цели были поставлены следующие **задачи**:

1. Произвести обзор и систематизацию современных публикаций о молекулярных механизмах болезни Альцгеймера и роли β -амилоида.
2. Проанализировать гипотезы происхождения пептида HАЕЕ в организме человека.
3. Провести биоинформатический поиск геномных последовательностей, потенциально кодирующих данный пептид в геноме человека.
4. Сделать сравнительный анализ полученных данных.

Научная новизна. Впервые реализован de novo вычислительный подход для поиска последовательностей исследуемого мотива НАЕЕ в геноме человека с использованием поиска вырожденных нуклеотидных последовательностей. В ходе исследования впервые была проведена аннотация найденных последовательностей с использованием данных GENCODE, что позволило выявить локализацию мотивов как в кодирующих областях, так и в некодирующих, включая альтернативные транскрипты и участки, потенциально ассоциированные с открытыми рамками считывания.

Практическая значимость. Данное исследование имеет важное значение в изучении потенциальных источников микропептидов, ассоциированных с болезнью Альцгеймера и другими нейродегенеративными заболеваниями. Полученные данные могут быть использованы для дальнейшего применения в методах машинного обучения, рибосомном профайлинге и протеономных подходах при исследовании тетрапептида НАЕЕ в организме человека.

1. ОБЗОР ЛИТЕРАТУРЫ

1.1. Болезнь Альцгеймера

Болезнь Альцгеймера (БА) является наиболее распространенной формой нейродегенеративных заболеваний и основной причиной деменции у лиц не только пожилого возраста [1], но и более раннего [2]. По данным Всемирной организации здравоохранения, на долю данного заболевания приходится около 60-70% всех случаев деменции [3]. Заболевание характеризуется прогрессирующим снижением когнитивных функций: нарушением памяти, речи, ориентации и способности к обучению, что со временем приводит к выраженной социальной и функциональной дезадаптации пациентов [4].

Патологические изменения при болезни Альцгеймера развиваются постепенно и могут начинаться за десятилетия до появления клинических симптомов [5]. Основными морфологическими признаками заболевания являются образование внеклеточных амилоидных бляшек и внутриклеточных нейрофибриллярных клубков [6]. Амилоидные бляшки преимущественно состоят из β -амилоидного пептида ($A\beta$), тогда как нейрофибриллярные клубки формируются из гиперфосфорилированного тау-белка [7].

Накопление β -амилоидного пептида рассматривается как один из ключевых факторов, запускающих каскад патологических процессов в нейронах [8]. Среди них выделяют нарушение синаптической передачи, активацию нейровоспалительных процессов, усиление окислительного стресса и нарушение внутриклеточного кальциевого гомеостаза [9]. В совокупности данные процессы приводят к постепенной гибели нейронов и прогрессирующей атрофии различных областей головного мозга, прежде всего гиппокампа и коры больших полушарий [10].

Несмотря на значительный прогресс в изучении заболевания, точные механизмы его развития остаются предметом активных исследований. В настоящее время рассматривается несколько патогенетических гипотез, среди

которых наиболее распространенной является амилоидная гипотеза, связывающая развитие заболевания с нарушением метаболизма β -амилоидного пептида и его последующей агрегацией [11].

1.2. Молекулярные основы болезни Альцгеймера

Одним из центральных элементов патогенеза болезни Альцгеймера является накопление β -амилоидного пептида ($A\beta$) в тканях головного мозга [12,13]. Считается, что избыточное образование и последующая агрегация данного пептида приводят к формированию внеклеточных амилоидных бляшек - одного из основных морфологических признаков заболевания [14]. Амилоидные бляшки преимущественно локализуются в коре больших полушарий и гиппокампе - областях мозга, играющих ключевую роль в процессах памяти и обучения [6].

Наряду с амилоидными бляшками важным патоморфологическим признаком болезни Альцгеймера является образование внутриклеточных нейрофибриллярных клубков, состоящих из гиперфосфорилированного тау-белка. В норме тау-белок выполняет функцию стабилизации микротрубочек, обеспечивая поддержание структуры нейронов и участие в аксональном транспорте. Однако при патологических условиях происходит гиперфосфорилирование данного белка, что приводит к нарушению его способности связываться с микротрубочками. В результате микротрубочковый цитоскелет дестабилизируется, нарушается транспорт молекул и органелл внутри клетки, что способствует развитию нейродегенеративных процессов [7].

Накопление β -амилоида и патологические изменения тау-белка сопровождаются нарушением синаптической передачи [15]. Установлено, что растворимые олигомерные формы β -амилоида способны взаимодействовать с рецепторами на поверхности нейронов и изменять работу синапсов. Это приводит к снижению эффективности передачи нервных импульсов и нарушению процессов долговременной потенциации, которая играет важную роль в механизмах формирования памяти. В результате постепенно развивается

синаптическая дисфункция, считающаяся одним из ранних событий в патогенезе болезни Альцгеймера [16].

Важную роль в развитии заболевания играют также процессы нейровоспаления. В ответ на накопление амилоидных агрегатов активируются клетки врожденного иммунитета центральной нервной системы - микроглия и астроциты. Активация микроглии сопровождается выделением провоспалительных цитокинов, хемокинов и реактивных форм кислорода. С одной стороны, данные процессы направлены на удаление патологических белковых агрегатов, однако при хронической активации они могут приводить к повреждению нейронов и усилению нейродегенерации [17].

Дополнительным фактором повреждения нейронов является усиление окислительного стресса. При болезни Альцгеймера наблюдается повышенное образование активных форм кислорода и нарушение работы антиоксидантных систем клетки. Окислительное повреждение может затрагивать липиды клеточных мембран, белки и нуклеиновые кислоты, что приводит к нарушению функций нейронов и их последующей гибели [9].

Существенную роль в патогенезе заболевания играют нарушения кальциевого гомеостаза. β -амилоид способен изменять проницаемость клеточных мембран и влиять на функционирование кальциевых каналов, что приводит к повышению внутриклеточной концентрации ионов кальция. Избыточное накопление кальция в нейронах может активировать ферменты, повреждающие клеточные структуры, а также инициировать процессы апоптоза [18].

Кроме того, при болезни Альцгеймера часто наблюдается дисфункция митохондрий. Митохондрии являются основными источниками энергии в клетке, и нарушение их работы приводит к снижению продукции АТФ, усилению образования активных форм кислорода и запуску механизмов клеточной гибели. Показано, что β -амилоид может непосредственно взаимодействовать с митохондриальными белками, нарушая их нормальное функционирование [19,20].

Таким образом, патогенез болезни Альцгеймера представляет собой сложный многофакторный процесс, включающий взаимодействие различных

молекулярных механизмов. Накопление β -амилоидного пептида, патологические изменения тау-белка, нейровоспаление, окислительный стресс и митохондриальная дисфункция в совокупности приводят к прогрессирующему нарушению функций нейронов и развитию нейродегенерации. В настоящее время одной из наиболее распространенных концепций, объясняющих развитие заболевания, остается амилоидная гипотеза, согласно которой именно нарушение метаболизма β -амилоида является ключевым событием, запускающим каскад патологических процессов [11].

1.3. Генезис β -амилоидного пептида

β -амилоидный пептид ($A\beta$) образуется в результате протеолитического процессинга амилоидного белка-предшественника (APP - amyloid precursor protein) [21,15]. APP представляет собой трансмембранный гликопротеин, широко экспрессируемый в различных тканях организма, включая нейроны центральной нервной системы [22]. Физиологическая роль APP полностью не установлена, однако предполагается его участие в регуляции нейрональной пластичности, клеточной адгезии и процессах роста нейронов [23]. Процессинг APP может происходить по двум основным путям. В первом случае - APP расщепляется α -секретазой в пределах участка, соответствующего последовательности β -амилоида. В результате образуется растворимый фрагмент sAPP α и мембранный фрагмент C83. Последующее расщепление C83 γ -секретазой приводит к образованию внутриклеточного домена APP (AICD) и короткого фрагмента, называемого P3, не обладающих выраженной нейротоксичностью (рис. 1). Данный путь считается физиологическим и предотвращает образование β -амилоидного пептида [24,25].

Альтернативный амилоидогенный путь начинается с действия β -секретазы (BACE1), которая расщепляет APP с образованием растворимого фрагмента sAPP β и мембранного фрагмента C99 (рис. 1). Далее C99 подвергается протеолизу γ -секретазным комплексом, состоящим из нескольких белков, включая

пресенилины, что приводит к образованию β -амилоидных пептидов различной длины [26,27]. Наиболее изученными являются формы $A\beta_{40}$ и $A\beta_{42}$, различающиеся длиной С-концевого участка [15]. Пептид $A\beta_{42}$ обладает более выраженной склонностью к агрегации и считается наиболее патогенной формой β -амилоида. Накопление $A\beta_{42}$ способствует формированию олигомерных структур, которые могут далее агрегировать в фибриллы и образовывать амилоидные бляшки [28].

В последние годы доказано, что наиболее токсичными являются не столько зрелые амилоидные фибриллы, сколько растворимые олигомерные формы $A\beta$, способные нарушать синаптическую передачу, изменять структуру клеточных мембран и индуцировать нейротоксические процессы [16]. В связи с этим значительное внимание уделяется поиску молекул, способных препятствовать агрегации β -амилоида или модифицировать процесс формирования амилоидных агрегатов.

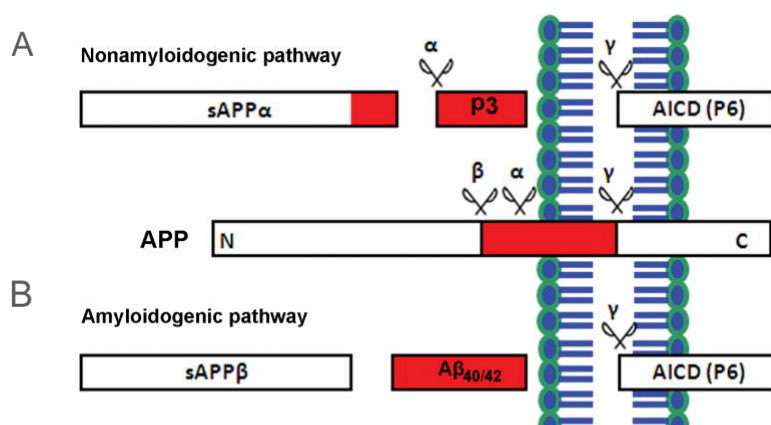


Рисунок 1. Схематическое изображение путей обработки и продуктов расщепления белка-предшественника амилоида (APP): А - неамилоидогенный путь обработки APP; В - амилоидогенный путь обработки APP. Иллюстрация взята из базы данных PubMed (<https://pmc.ncbi.nlm.nih.gov/articles/PMC5296245/>).

1.4. История разработки пептида НАЕЕ

В рамках поиска соединений, способных ингибировать агрегацию β -амилоидного пептида, значительное внимание уделяется коротким пептидам. Такие молекулы могут взаимодействовать с β -амилоидом и нарушать процессы формирования агрегатов за счет конкурентного связывания или изменения конформации белка [29].

Одним из таких соединений является тетрапептид НАЕЕ (His-Ala-Glu-Glu) [30]. Данный пептид был предложен в качестве потенциального ингибитора агрегации β -амилоидного пептида. Предполагалось, что его короткая последовательность может взаимодействовать с определенными участками β -амилоида и препятствовать образованию фибриллярных структур [31].

Экспериментальные исследования показали, что НАЕЕ способен влиять на процессы агрегации β -амилоидного пептида *in vitro* [32]. В ряде работ было продемонстрировано снижение скорости образования амилоидных фибрилл в присутствии данного пептида. Кроме того, в экспериментальных моделях болезни Альцгеймера обсуждалась возможность его нейропротекторного действия [33].

Интерес к подобным коротким пептидам связан также с их потенциальными преимуществами в качестве терапевтических агентов. К ним относятся относительно небольшая молекулярная масса, возможность синтетического получения и потенциально высокая специфичность взаимодействия с белковыми мишенями [34].

Несмотря на ограниченное количество исследований, посвященных непосредственно НАЕЕ, данный пептид рассматривается как один из представителей коротких регуляторных пептидов, способных модифицировать процессы агрегации β -амилоида [35].

1.5. Открытие эндогенности НАЕЕ

Первоначально НАЕЕ рассматривался исключительно как синтетический пептид, разработанный для изучения механизмов ингибирования агрегации β -амилоида [32,36,31]. Однако в дальнейшем появились данные, указывающие на возможность существования коротких пептидов аналогичной длины в составе так называемого эндогенного пептидома организма [37,38].

Современные методы протеомного анализа, включая масс-спектрометрию, позволяют выявлять все большее количество коротких пептидов, образующихся в результате внутриклеточного протеолиза белков [39]. Такие пептиды могут возникать в процессе деградации белков протеосомой [40,41], лизосомальными ферментами или другими протеолитическими системами клетки.

В последние годы было доказано, что многие короткие пептиды длиной от трех до десяти аминокислот присутствуют в различных тканях организма [42] и могут выполнять регуляторные функции [43]. Они способны участвовать в межклеточной сигнализации, регуляции экспрессии генов и модуляции активности ферментов [44].

В связи с этим была выдвинута гипотеза о том, что некоторые короткие пептиды, ранее рассматривавшиеся исключительно как синтетические молекулы, могут иметь эндогенные аналоги, образующиеся в организме в результате протеолитических процессов.

Если пептид НАЕЕ действительно может образовываться *in vivo*, возникает вопрос о возможных источниках его образования и механизмах генерации данного пептида в клетках организма.

1.6. Гипотезы источника НАЕЕ в организме человека

Существует несколько гипотез относительно возможного происхождения коротких пептидов в организме человека, включая пептиды, содержащие последовательность НАЕЕ.

Одной из возможных гипотез является образование таких пептидов в результате протеолитической деградации клеточных белков. В клетках постоянно происходит обновление белков, сопровождающееся их расщеплением протеосомными и лизосомальными ферментами. В процессе такого протеолиза могут образовываться короткие олигопептиды различной длины, включая тетрапептиды [45].

Другим возможным источником является альтернативный процессинг белков-предшественников, при котором образуются короткие пептидные фрагменты с потенциальной биологической активностью. Подобные механизмы известны для ряда нейропептидов и сигнальных молекул [46].

Также обсуждается возможность существования коротких открытых рамок считывания (sORF) в геноме человека. Такие рамки считывания могут кодировать короткие пептиды, ранее не аннотированные в составе классических белок-кодирующих генов. В последние годы благодаря развитию методов геномики и протеомики было обнаружено значительное количество подобных коротких пептидов [47,48].

Кроме того, потенциальным источником коротких пептидов могут являться альтернативные транскрипты генов, включая бицистронные мРНК и различные варианты сплайсинга. Такие транскрипты могут содержать дополнительные короткие рамки считывания, приводящие к синтезу небольших пептидных продуктов [49].

1.7. Постановка задачи

Проверка гипотез о возможном геномном происхождении пептида НАЕЕ требует анализа нуклеотидных последовательностей генома человека и поиска участков, потенциально кодирующих данный аминокислотный мотив. Современные методы биоинформатики позволяют проводить подобный поиск с использованием алгоритмов анализа геномных данных и сопоставления найденных последовательностей с аннотацией генов [50].

В связи с этим одним из возможных подходов к изучению происхождения НАЕЕ является вычислительный анализ генома человека, направленный на выявление нуклеотидных последовательностей, способных кодировать данный пептид, и определение их локализации в белок-кодирующих участках генома.

Поскольку пептид НАЕЕ имеет две посттрансляционные модификации, то возможный источник эндогенного НАЕЕ может быть предложен на основании следующих соображений: большинство физиологически активных пептидов, включая Аβ, образуются из белков-предшественников под действием определенных протеаз и/или пептидаз пептидного ряда. Многие эндогенные пептиды подвергаются ферментативному ацетилированию на N-конце. Ацетилирование является одной из основных посттрансляционных модификаций белка в клетке и оказывает многообразное воздействие, как на уровень белка, так и на уровень метаболома. Ацетильная группа поставляется ацетил-коферментом А и может быть посттрансляционно присоединена либо к α-аминогруппе N-конца белков, либо к ε-аминогруппе остатков лизина. Эти реакции катализируются различными N-концевыми и лизин ацетилтрансферазами 3 (КАТ3). Но большинство из них не могут модифицировать пептиды с первым остатком гистидина. C9JB11, Q8N8M0, Q8N8M0-2 являются потенциальными кандидатами для ацетилирования N-конца пептида НАЕЕG.

Напротив, количество эндогенных пептидов, амидированных на C-конце, очень ограничено, и C-амидирование происходит за счет ферментативного превращения остатка глицина на C-конце в амидную группу [51]. Поскольку эндогенный НАЕЕ содержит как N-ацетильную, так и C-амидную группы, белок-предшественник должен содержать пентапептидный фрагмент НАЕЕG в своей аминокислотной последовательности.

α-амидирование является критическим фактором, определяющим биологическую активность многих нейропептидов. Окислительное расщепление катализируется пептидилглицин-α-амидирующей монооксигеназой (РАМ), ферментом, локализованным в транс-сети аппарата Гольджи и секреторных гранулах нервной и эндокринной тканей (рис. 2). РАМ является

бифункциональным ферментом с двумя доменами: пептидилглицин- α -гидроксилирующей монооксигеназой и пептидил- α -гидроксиглицин- α -амидирующей лиазой (рис. 2). Эти два домена действуют последовательно, образуя активный α -амидированный пептидный продукт и глиоксилат (рис. 3).

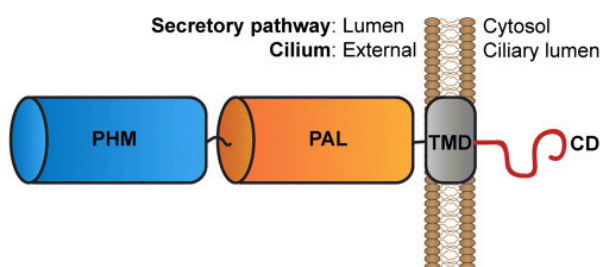


Рисунок 2. Доменная организация PAM. Два каталитических ядра (PHM и PAL) расположены в просвете секреторного пути и на поверхности ресничек. За ними следует трансмембранный регион (TMD) и неструктурированный цитозольный домен (CD), содержащий сигналы, необходимые для транспорта по биосинтетическому и эндоцитозному путям. Иллюстрация взята из базы данных PubMed (<https://pmc.ncbi.nlm.nih.gov/articles/PMC6529398/figure/Fig4/>).

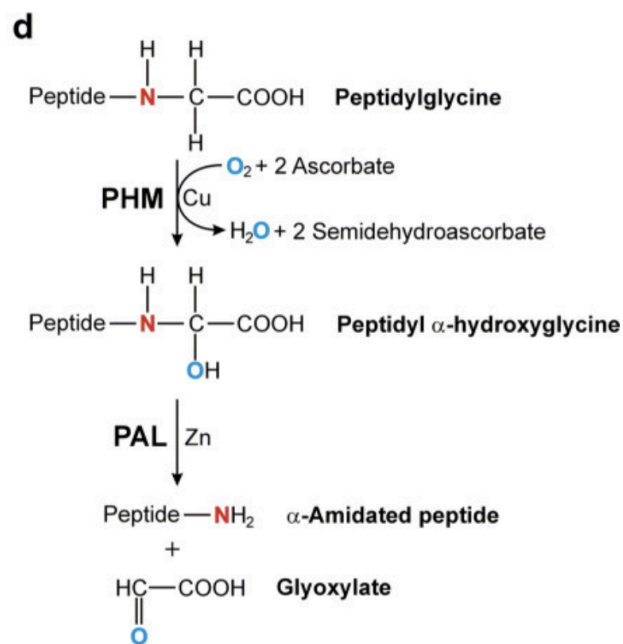


Рисунок 3. Механизм реакции амидирования. Пептидные предшественники с удлиненным глицином стереоспецифически гидроксилируются на атоме углерода C_α с помощью PHM в процессе, требующем меди, связанной в двух разных местах, молекулярного кислорода и двух молекул аскорбата. Впоследствии связь $\text{C}_\alpha\text{-N}$ расщепляется PAL, образуя амидированный пептид и высвобождая глиоксилат. Иллюстрация взята из базы данных PubMed (<https://pmc.ncbi.nlm.nih.gov/articles/PMC6529398/figure/Fig4/>).

2. МАТЕРИАЛЫ И МЕТОДЫ

2.1. Материалы

В данной работе использовалась референсная сборка генома человека GRCh 38 в формате FASTA, первичная сборка, а также аннотация генов GENCODE v49 в формате GTF. Данные были взяты из базы данных GENCODE.

В качестве исследуемой последовательности использовался вырожденный нуклеотидный мотив CAYGCNGARGARGGN, соответствующий аминокислотному мотиву исследуемого пептида - HAEEG, и обратная комплементарная ему последовательность - NCCYTCYTCNGCRTG. Такой выбор последовательности позволил повысить специфичность поиска и снизить вероятность выявления случайных совпадений в геноме человека.

Анализ последовательностей, представленный в базе данных UniProt (<https://www.uniprot.org/>), продемонстрировал сохранение остатка глицина в исследуемом мотиве у различных белков человека, что позволило рассматривать HAEEG как более биологически значимый и консервативный мотив.

Для функциональной аннотации идентифицированных генов-кандидатов из CDS-регионов использовались базы данных UniProt (UniProt Consortium, 2025) и Human Protein Atlas (v23). Для анализа тканевой экспрессии привлекались данные проекта GTEx (v8).

2.2. Методы

2.2.1. Фильтрация и аннотирование найденных последовательностей

Для поиска локализации исследуемого мотива был разработан алгоритм с использованием регуляторных выражений для поиска вырожденной нуклеотидной последовательности CAYGCNGARGARGGN для прямой цепи и NCCYTCYTCNGCRTG для обратной с учетом нотации IUPAC (рис. 4).

Таким образом, общее число последовательностей для анализа включало 128 варианта последовательностей для одной цепи и 256 последовательностей для обеих цепей. Анализ проводился как для прямой цепи ДНК, так и для обратной последовательности. Учет обеих цепей позволил повысить полноту поиска потенциальных последовательностей, соответствующих исследуемому мотиву.

Для каждой найденной позиции мотива определялись следующие параметры: хромосома, стартовая и конечная координаты, цепь ДНК (+/-). Затем проводилось сопоставление с GTF-аннотацией для определения принадлежности мотива к одному из следующих типов регионов: CDS (кодирующая последовательность), UTR (нетранслируемая область), межгенный регион (отсутствие пересечения с аннотированными генами ± 2 кб).

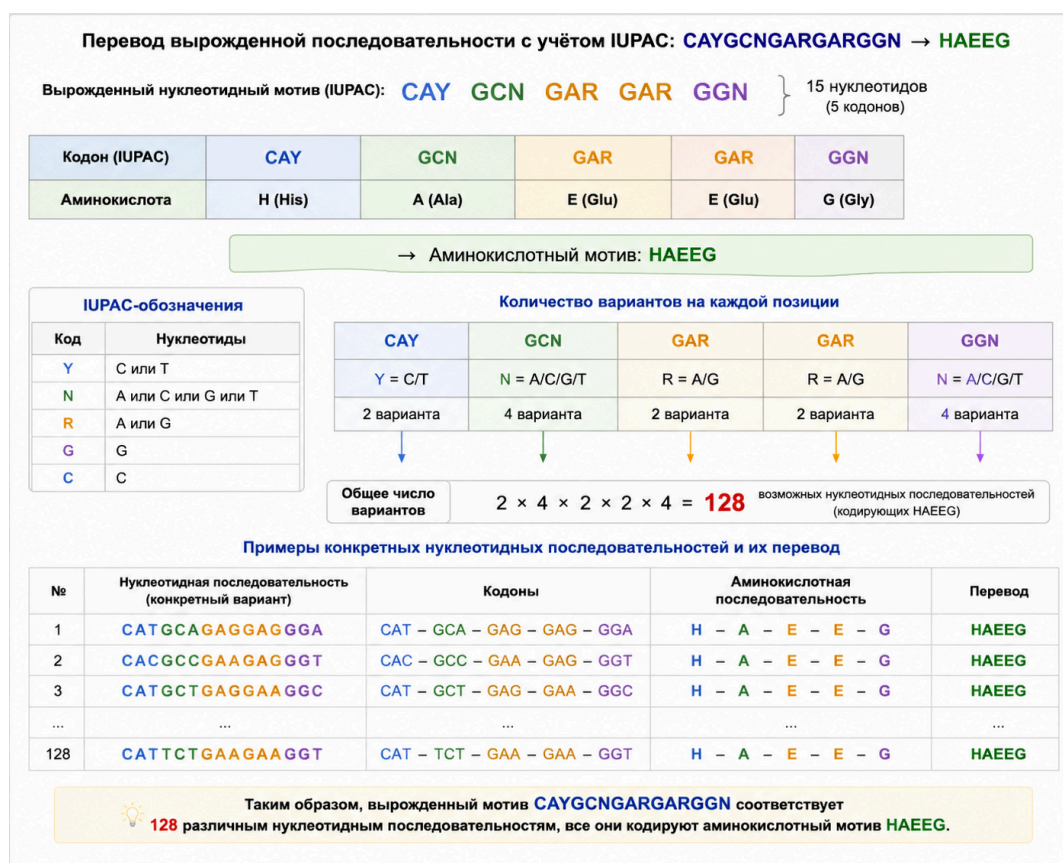


Рисунок 4. Перевод вырожденной последовательности CAYGCNGARGARGGN с учетом нотации IUPAC.

2.2.2. Идентификация открытых рамок считывания в межгенных и 5UTR регионах

Для межгенных и 5UTR регионов, содержащих мотив, проводился поиск открытых рамок считывания (ORF). Из геномной последовательности извлекался регион размером ± 300 нуклеотидов относительно центра мотива. Для извлеченной последовательности анализировались все 6 рамок считывания (3 на прямой цепи и 3 на комплементарной). ORF считалась валидной при наличии старт-кодона (ATG, CTG, TTG или GTG) и стоп-кодона (TAA, TAG, TGA) в одной рамке, а также при условии, что мотив НАЕЕ полностью находится внутри предсказанного пептида. Трансляция нуклеотидной последовательности в пептидную выполнялась с использованием стандартного генетического кода.

2.2.3. Статистический анализ

Статистическая обработка данных выполнялась с использованием библиотеки pandas (v2.1.0). Визуализация распределений длин пептидов, типов старт-кодонов и локализации мотива НАЕЕ относительно концов пептида выполнялась с помощью библиотеки matplotlib (v3.8.0) и seaborn (v0.13.0). Кластеризация генов по функциональным категориям проводилась на основе аннотаций Gene Ontology (GO), полученных через UniProt.

2.2.4. Биоинформатическое программное обеспечение и базы данных

Алгоритмы анализа и аннотирования найденных последовательностей был выполнен на языке Python 3.12.7. Для обработки геномных данных применялись библиотеки Biopython и pandas (v2.1.0).

2.2.5. Валидация результатов с помощью BLASTN

Для контроля специфичности поиска и исключения ложноположительных находок проводилась валидация с использованием сервиса BLASTN (NCBI, <https://blast.ncbi.nlm.nih.gov>). Для каждого уникального локуса, содержащего мотив, выполнялся BLAST-поиск против референсного генома человека GRCh38.p14.

Критериями валидации служили:

- 1) полное совпадение с целевым локусом в геноме;
- 2) отсутствие множественных неспецифических совпадений (менее 3 высокоточных хитов с $e\text{-value} < 10^{-5}$).

3. РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

3.1. Поиск вырожденного мотива в геноме человека

С целью выявления потенциальных участков кодирования исследуемого белка HAEE был проведен поиск вырожденной нуклеотидной последовательности CAYGCNGARGARGGN для прямой цепи и NCCYTCYTCNGCRTG для обратной в геноме человека. В результате было выявлено 654 (рис. 5) вхождения исследуемого мотива в геноме человека. Из них 307 последовательностей были обнаружены в прямой цепи ДНК и 347 - в комплементарной.

```
results > motif_hits_both_strands.csv
25 chr1,214835895,214835910,CATGCAGAAGAAGGG,+
26 chr1,219975266,219975281,CACGCAGAGGAGGGG,+
27 chr1,224656156,224656171,CATGCTGAGGAAGGT,+
28 chr1,225398266,225398281,CATGCTGAGGAAGGG,+
29 chr1,231698402,231698417,CATGCAGAGGAGGGG,+
30 chr1,247268229,247268244,CATGCTGAAGAAGGA,+
31 chr1,247595490,247595505,CATGCAGAAGAAGGA,+
32 chr1,3361245,3361260,ACCCTCCTCAGCATG,-
33 chr1,5389438,5389453,ACCCTCTTCAGCATG,-
34 chr1,7520860,7520875,GCCTTCCTCAGCATG,-
35 chr1,9391061,9391076,CCCCTCCTCAGCATG,-
36 chr1,13534894,13534909,TCCTTCCTCAGCATG,-
37 chr1,14439894,14439909,TCCTTCTTCAGCATG,-
38 chr1,26085000,26085015,ACCCTCTTCTGCATG,-
39 chr1,30860007,30860022,TCCCTCCTCTGCATG,-
40 chr1,31783817,31783832,GCCCTCTTCTGCATG,-
```

Рисунок 5. Локализация мотива в геноме человека на прямой и обратной цепи ДНК.

Из результатов видно, что мотив пептида НАЕЕ широко распространен как внутри одной хромосомы, так и большим разнообразием расположения на других хромосомах человека (табл. 1).

Таблица 1. Распространение мотива пептида НАЕЕ на хромосомах

Хромосома	Количество найденных мотивов	Цепь ДНК (+/-)	Хромосома	Количество найденных мотивов	Цепь ДНК (+/-)
chr 1	30	+	chr 10	21	-
chr 1	31	-	chr 11	17	+
chr 2	19	+	chr 11	16	-
chr 2	26	-	chr 12	12	+
chr 3	14	+	chr 12	11	-
chr 3	20	-	chr 13	8	+
chr 4	16	+	chr 13	10	-
chr 4	22	-	chr 14	8	+
chr 5	22	+	chr 14	12	-
chr 5	20	-	chr 15	8	+
chr 6	16	+	chr 15	12	-
chr 6	20	-	chr 16	10	+
chr 7	13	+	chr 16	9	-
chr 7	20	-	chr 17	9	+
chr 8	20	+	chr 17	14	-
chr 8	13	-	chr 18	12	+
chr 9	14	+	chr 18	11	-
chr 9	19	-	chr 19	5	+
chr 10	22	+	chr 19	7	-

Продолжение таблицы 1.

Хромосома	Количество найденных мотивов	Цепь ДНК (+/-)
chr 20	7	+
chr 20	8	-
chr 21	4	+
chr 21	2	-
chr 22	11	+
chr 22	6	-
chr X	10	+
chr X	17	-

Наличие большого количества найденных локусов может свидетельствовать о потенциальной функциональной значимости исследуемой последовательности, а также о возможности ее участия в регуляции или структурной организации кодирующих участков генома. Это предположение указывает на необходимость дальнейшего анализа локализации мотива относительно аннотированных генов и кодирующих участков.

3.1.1. Аннотирование и фильтрация найденных последовательностей

Для определения функциональной принадлежности найденных последовательностей проводилось сопоставление координат выявленных мотивов с аннотацией генома человека в формате GTF.

Для начала определялись участки пересечения найденных мотивов с аннотированными CDS-регионами белок-кодирующих генов. Таким образом,

проведенное аннотирование позволило оставить лишь малую часть из выявленных локусов (табл. 2).

Таблица 2. Результат поиска пересечения найденных мотивов с аннотированными CDS-регионами белок-кодирующих генов

Хромосома	Участок	Количество изоформ	Цепь ДНК (+/-)
chr 1	6587227 - 6587242	15	+
	171103787 - 171103802	11	+
	204201667 - 204201682	2	+
	110389309 - 110389324	11	-
	209675783 - 209675798	4	-
chr 2	98396594 - 98396609	5	+
	218364699 - 218364714	6	+
	100417015 - 100417030	1	-
chr 3	9867265 - 9867280	16	+
	32452409 - 32452424	7	-
	45964371 - 45964386	5	-
chr 4	960701 - 960716	7	-
chr 5	16616932 - 16616947	5	-
chr 6	170289676 - 170289691	11	+
	136592257 - 136592272	9	-
chr 7	149476965 - 149476980	5	+
chr 9	137957641 - 137957656	12	+
	131446509 - 131446524	3	-
chr 10	133568407 - 133568422	1	+
chr 11	17395867 - 17395882	15	+

Продолжение таблицы 2.

	34164762 - 34164777	2	+
	56576947 - 56576962	1	+
	56612728 - 56612743	1	+
	65346373 - 65346388	1	-
	72013174 - 72013189	25	-
chr 12	6909897 - 6909912	5	+
	3696835 - 3696850	9	-
chr 15	86988066 - 86988081	1	-
chr 16	1338943 - 1338958	20	+
	90037245 - 90037260	8	+
	68361623 - 68361638	17	-
chr 17	82054587 - 82054602	46	+
	44401098 - 44401113	4	-
	81976916 - 81976931	1	-
chr 19	35733260 - 35733275	5	-
	45779442 - 45779457	19	-
chr 22	29825107 - 29825122	57	+
	30533993 - 30534008	5	-

Анализ GTF-аннотации показал, что часть найденных мотивов соответствует нескольким транскрипционным изоформам одного и того же гена. Это связано, скорее всего, с альтернативным сплайсингом и существованием различных вариантов транскриптов, имеющих различную структуру CDS-регионов.

Для дальнейшего анализа из результатов были исключены последовательности, относящиеся к некодирующим транскриптам типа “nonsense

mediated decay”, некодирующие и межгенные участки и оставлены только последовательности, относящиеся к категории “белок-кодирующий регион”.

Таким образом, проведенное аннотирование позволило оставить последовательности, которые являются наиболее перспективными для дальнейшего анализа (табл. 3).

Из данных видно, что исследуемый мотив распределен по геному неравномерно. Наибольшее количество изоформ последовательностей находятся на 17 и 22 хромосоме человека.

Таблица 3. Результат сортировки последовательностей, относящихся к категории “белок-кодирующий регион”

Хромосома	Участок мотива	Количество изоформ	Цепь ДНК (+/-)
chr 1	6587227 - 6587242	15	+
	171103787 - 171103802	11	+
	204201667 - 204201682	2	+
	110389309 - 110389324	8	-
	209675783 - 209675798	4	-
chr 2	98396594 - 98396609	5	+
	218364699 - 218364714	6	+
	100417015 - 100417030	1	-
chr 3	9867265 - 9867280	16	+
	32452409 - 32452424	6	-
	45964371 - 45964386	5	-
chr 4	960701 - 960716	7	-
chr 5	16616932 - 16616947	4	-
chr 6	170289676 - 170289691	11	+

Продолжение таблицы 3.

	136592257 - 136592272	9	-
chr 7	149476965 - 149476980	5	+
chr 9	137957641 - 137957656	9	+
	131446509 - 131446524	3	-
chr 10	133568407 - 133568422	1	+
chr 11	17395867 - 17395882	11	+
	34164762 - 34164777	2	+
	56576947 - 56576962	1	+
	56612728 - 56612743	1	+
	72013174 - 72013189	24	-
chr 12	6909897 - 6909912	5	+
	3696835 - 3696850	9	-
chr 15	86988066 - 86988081	1	-
chr 16	1338943 - 1338958	17	+
	90037245 - 90037260	8	+
	68361623 - 68361638	17	-
chr 17	82054587 - 82054602	46	+
	44401098 - 44401113	4	-
	81976916 - 81976931	1	-
chr 19	35733260 - 35733275	3	-
	45779442 - 45779457	19	-
chr 22	29825107 - 29825122	57	+
	30533993 - 30534008	5	-

В ходе дальнейшего исследования идентифицировано 19 уникальных генов для прямой цепи ДНК и 18 генов для обратной цепи, в пределах белок-кодирующих участков которых локализован исследуемый мотив. (табл. 4).

Таблица 4. Представленность исследуемого мотива пептида НАЕЕ в генах человека

Uniprot индекс	Последовательность	Количество изоформ	Протеин	Ген	Хромосома	Цепь ДНК (+/-)
Q8N9N2	CATGCGGA GGAAGGT	57	Activating signal cointegrator 1 complex subunit 1	ASCC2	chr 22	+
Q13098	CACGCGGA AGAAGGC	46	COP9 signalosome complex subunit 1	GPS1	chr 17	+
Q14980	GCCTTCTT CAGCATG	24	Nuclear mitotic apparatus protein 1	NUMA 1	chr 11	-
Q09013	CCCCTCTT CAGCATG	19	Myotonin-p rotein kinase	DMPK	chr 19	-
Q9NY59	CCCCTCCT CTGCATG	17	Sphingomy elin phosphodies terase 3	SMPD3	chr 16	-
O94812	CATGCTGA AGAAGGG	17	BAI1-associated protein 3	BAIAP 3	chr 16	+
Q96AQ7	CATGCTGA AGAGGGC	16	Lipid transferase CIDEC	CIDEC	chr 3	+

Продолжение таблицы 4.

P52824	GCCTTCCT CAGCATG	7	Diacylglycerol kinase theta	DGKQ	chr 4	-
P10074	CATGCAGA AGAAGGA	15	Zinc finger and BTB domain-containin g protein 48	ZBTB48	chr 1	+
P31513	CATGCAGA GGAGGGC	11	Flavin-containing monooxygenase 3	FMO3	chr 1	+
O00548	CACGCGGA AGAAGGT	11	Delta-like protein 1	DLL1	chr 6	+
Q09428	CATGCGGA AGAAGGC	11	ATP-binding cassette sub-family C member 8	ABCC8	chr 11	+
Q99683	TCCTTCCT CAGCATG	9	Mitogen-activated protein kinase kinase kinase 5	MAP3K 5	chr 6	-
Q9BSW2	GCCTTCCT CAGCATG	9	EF-hand calcium-binding domain-containin g protein 4B	CRACR 2A	chr 12	-
Q00975	CACGCAGA GGAGGGA	9	Voltage-dependen t N-type calcium channel subunit alpha-1B	CACNA 1B	chr 9	+
O15374	TCCCTCTT CAGCATG	8	Probable monocarboxylate transporter 5	SLC16A 4	chr 1	-
O95995	CATGCGGA AGAAGGA	8	Dynein regulatory complex subunit 4	GAS8	chr 16	+

Продолжение таблицы 4.

Q96FZ5	ACCTTCCT CTGCATG	6	CKLF-like MARVEL transmembrane domain-containi ng protein 7	CMTM 7	chr 3	-
Q7Z7H3	CACGCGGA GGAGGGC	6	Ciliogenesis-asso ciated TTC17-interactin g protein	CATIP	chr 2	+
Q53EV4	CATGCTGA AGAAGGT	5	Leucine-rich repeat-containing protein 23	LRRC2 3	chr 12	+
B5MCN 3	ACCTTCCT CAGCATG	5	SEC14-like protein 6	SEC14L 6	chr 22	-
Q6NUN 9	CATGCTGA AGAGGGA	5	Zinc finger protein 746	ZNF746	chr 7	+
Q9BQS8	TCCTTCCT CAGCATG	5	FYVE and coiled-coil domain-containi ng protein 1	FYCO1	chr 3	-
Q16281	CACGCTGA AGAAGGT	5	Cyclic nucleotide-gated channel alpha-3	CNGA3	chr 2	+
Q9H6L5	TCCCTCCT CGGCGTG	4	Reticulophagy regulator 1	RETRE G1	chr 5	-
Q9UKJ3	CCCTTCCT CCGCATG	4	G patch domain-containi ng protein 8	GPATC H8	chr 17	-
P27469	GCCCTCTT CGGCGTG	4	G0/G1 switch protein 2	G0S2	chr 1	-
Q5JSZ5	CCCCTCCT CGGCATG	3	Protein PRRC2B	PRRC2 B	chr 9	-
Q9UMN 6	CCCCTCCT CGGCGTG	3	Histone-lysine N-methyltransfer ase 2B	KMT2B	chr 19	-

Продолжение таблицы 4.

Q8N961	CATGCTGA GGAGGGG	2	Ankyrin repeat and BTB/POZ domain-contai ning protein 2	ABTB2	chr 11	+
Q6ZVE7	CATGCCGA GGAGGGG	2	Vesicle transport protein GOT1A	GOLT1 A	chr 1	+
Q6IEU7	CATGCAGA AGAGGGT	1	Olfactory receptor 5M10	OR5M1 0	chr 11	+
Q8NGP8	CATGCAGA AGAGGGT	1	Olfactory receptor 5M1	OR5M1	chr 11	+
Q96MI9	ACCTTCTT CTGCATG	1	Cytosolic carboxypeptida se 4	AGBL1	chr 15	-
O43529	GCCTTCCT CTGCATG	1	Carbohydrate sulfotransferas e 10	CHST1 0	chr 2	-
Q9BZE9	CCCCTCCT CAGCATG	1	Tether containing UBX domain for GLUT4	ASPSC R1	chr 17	-
Q8N0S2	CACGCAGA GGAAGGC	1	Synaptonemal complex central element protein 1	SYCE1	chr 10	+

Наиболее высокая представленность мотива в прямой цепи зафиксирована у генов ASCC2, GPS1 и BAIAP3. Что касается обратной цепи, рекордсменами по числу изоформ с мотивом выступают гены NUMA1 и DMPK.

Для оценки потенциальной способности идентифицированных генов служить источниками эндогенного пептида HАЕЕ был проведен анализ их тканевой экспрессии с использованием баз данных UniProt и Human Protein Atlas. Установлено, что среди выявленных 37 генов, содержащих мотив HАЕЕG в

кодирующих последовательностях, 10 характеризуются высокой экспрессией в тканях нервной системы, а 5 - в эндокринных органах. Это обстоятельство представляет особый интерес, поскольку фермент пептидилглицин- α -амидирующая монооксигеназа (РАМ), ответственный за С-амидирование пептидов, локализован преимущественно в секреторных гранулах нервных и эндокринных клеток. Исходя из этого, наиболее перспективными кандидатами на роль источника эндогенного НАЕЕ призваны следующие белки:

- 1) CACNA1B (субъединица α -1В потенциал-зависимого кальциевого канала N-типа) - экспрессируется исключительно в нейронах и играет ключевую роль в экзоцитозе нейромедиаторов;
- 2) VAIAP3 (VAI1-ассоциированный белок 3) - участвует в кальций-зависимом экзоцитозе в нейронах и эндокринных клетках;
- 3) ABCC8 (АТФ-связывающий кассетный белок подсемейства С) - высокоспецифичен для β -клеток поджелудочной железы, секретирующих инсулин.

Таким образом, проведенный анализ позволил сузить круг потенциальных источников эндогенного НАЕЕ до группы белков, экспрессирующихся в нервной и эндокринной системах, с выделением трех генов в качестве приоритетных кандидатов для дальнейшего экспериментального исследования.

3.1.2. Контрольная валидация результатов с использованием BLASTN

Для контрольной проверки корректности найденных последовательностей и исключения возможных артефактов алгоритма поиска была проведена валидация всех идентифицированных локусов с использованием сервиса BLASTN базы данных NCBI (<https://blast.ncbi.nlm.nih.gov>).

Результаты BLASTN-анализа подтвердили соответствие 33 исследуемых последовательностей из 37 по референсному геному человека GRCh38.p14. Для 33 локусов наблюдалась высокая степень идентичности (100% совпадение) с

аннотированными участками геном. Для 4 локусов (гены SYCE1, ASPSCR1, SEC14L6, GAS8) значимых совпадений с заявленными генами обнаружено не было, а потому они были исключены из списка потенциальных источников эндогенного пептида HАЕЕ, однако сохранены в таблице для полноты представления результатов (табл. 5).

Таблица 5. Результаты BLASTN-валидации для потенциальных кандидатов источника пептида HАЕЕ

Ген	Последовательность мотива	Длина и идентичность	e-value	Координаты совпадения
GPS1	CACGCGGAAGAA GGC	15/15 (100%)	6,4	chr17:82054587-82054602
ASCC2	CATGCGGAGGAA GGT	15/15 (100%)	6,4	chr22:29825107-29825122
BAIAP3	CATGCTGAAGAA GGG	15/15 (100%)	6,4	chr16:1338943-1338958
NUMA1	GCCTTCTTCAGCA TG	15/15 (100%)	6,4	chr11:72013174-72013189
SMPD3	CCCCTCCTCTGCA TG	15/15 (100%)	6,4	chr16:68361623-68361638
DMPK	CCCCTCTTCAGCA TG	15/15 (100%)	6,4	chr19:45779442-45779457

Примечание: полный список результатов для всех 37 генов представлен в Приложении А.

Полученные совпадения соответствовали ранее определенным координатам мотивов и подтверждали корректность их локализации в соответствующих хромосомных регионах.

Полученные данные свидетельствуют о том, что большинство значимых выявленных локусов действительно присутствуют в геноме человека и не являются артефактами алгоритма поиска или обработки последовательностей.

3.1.3. Анализ расположения мотивов относительно сплайсируемых участков

Для оценки возможной связи исследуемого мотива с процессами сплайсинга был проведен анализ расположения найденных последовательностей относительно границ экзонов.

Последовательности классифицировались как расположенные вблизи сплайсируемых участков при нахождении в пределах 20 нуклеотидов от границ экзонных участков, определенных на основании GTF-аннотации генома человека.

Анализ показал, что значительная часть мотивов локализуется вблизи границ экзонов (1-7 нуклеотидов) - 11 мотивов. 9 мотивов находится неподалеку от границ (10-35 нуклеотидов). 12 мотивов находится на значительном расстоянии (38-735 нуклеотидов) от границ сплайсируемых участков (табл. 6).

Таблица 6. Расположение исследуемых мотивов относительно сплайсируемых участков

Хромосома	Участок	Ген	Степень близости от сплайсируемого участка (п.н.)	Цепь ДНК (+/-)
chr 22	29825107 - 29825122	ASCC2	1	+
chr 17	82054587 - 82054602	GPS1	1	+
chr 11	72013174 - 72013189	NUMA1	123	-
chr 19	45779442 - 45779457	DMPK	3	-
chr 16	68361623 - 68361638	SMPD3	20	-
chr 16	1338943 - 1338958	BAIAP3	31	+
chr 3	9867265 - 9867280	CIDEC	16	+
chr 1	6587227 - 6587242	ZBTB48	1	+
chr 1	171103787 - 171103802	FMO3	2	+
chr 6	170289676 - 170289691	DLL1	10	+

Продолжение таблицы 6.

chr 11	17395867 - 17395882	ABCC8	3	+
chr 6	136592257 - 136592272	MAP3K5	54	-
chr 12	3696835 - 3696850	CRACR2A	63	-
chr9	137957641 - 137957656	CACNA1B	31	+
chr1	110389309 - 110389324	SLC16A4	31	-
chr 4	960701 - 960716	DGKQ	5	-
chr 3	32452409 - 32452424	CMTM7	17	-
chr 2	218364699 - 218364714	CATIP	38	+
chr 12	6909897 - 6909912	LRRC23	7	+
chr 7	149476965 - 149476980	ZNF746	43	+
chr 3	45964371 - 45964386	FYCO1	35	-
chr 2	98396594 - 98396609	CNGA3	735	+
chr 5	16616932 - 16616947	RETREG1	21	-
chr 17	44401098 - 44401113	GPATCH8	224	-
chr 1	209675783 - 209675798	G0S2	130	-
chr 9	131446509 - 131446524	PRRC2B	4	-
chr 19	35733260 - 35733275	KMT2B	210	-
chr 11	34164762 - 34164777	ABTB2	44	+
chr 1	204201667 - 204201682	GOLT1A	19	+
chr 11	56576947 - 56576962	OR5M10	173	+
chr 11	56612728 - 56612743	OR5M1	776	+
chr 15	86988066 - 86988081	AGBL1	7	-
chr 2	100417015 - 100417030	CHST10	4	-

Особый интерес представляет анализ расположения мотива HAEEG относительно границ экзонов у трех наиболее перспективных генов-кандидатов,

выделенных на основе тканевой экспрессии и функциональной значимости: *CACNA1B*, *BAIAP3* и *ABCC8*.

Для гена *ABCC8* мотив расположен на расстоянии 3 нуклеотидов от границы экзона, что попадает в категорию «непосредственная близость» (1-10 п.н.). Такое расположение предполагает, что мотив находится вблизи сплайс-сайта и его включение в конечный транскрипт может зависеть от альтернативного сплайсинга. Учитывая, что *ABCC8* экспрессируется преимущественно в β -клетках поджелудочной железы, можно предположить, что продукция пептида НАЕЕ из этого белка может быть тканеспецифичной и регулируемой на уровне сплайсинга.

Для генов *CACNA1B* и *BAIAP3* мотив расположен на расстоянии 31 нуклеотида от границы экзона, что относится к категории «умеренное удаление» (11-50 п.н.). Такое расположение не исключает возможности альтернативного сплайсинга, но с меньшей вероятностью указывает на прямую зависимость от него.

Остальные полученные результаты могут свидетельствовать о потенциальной связи исследуемого мотива с участками, вовлеченными в процессы сплайсинга и представляют интерес с точки зрения функциональной роли данных последовательностей, несмотря на отсутствие прямых экспериментальных подтверждений.

3.2. Анализ дополнительных геномных областей, содержащих мотив пептида НАЕЕ

Помимо CDS-регионов, исследуемый мотив был найден в некодирующих областях генома человека (табл. 7). Особый интерес представили участки, найденные в UTR-регионах, вблизи стартовых кодонов, а также в межгенных регионах, поскольку подобные области потенциально могут содержать короткие открытые рамки считывания (sORFs) и кодировать функциональные микропротеины.

Таблица 7. Анализ части мотивов НАЕЕ, найденных в некодирующих регионах

Хромосома	Участок	Участок генома	Количество изоформ мотива	Цепь ДНК (+/-)
chr 1	51300371 - 51300386	UTR	1	+
	171103787 - 171103802	UTR	1	+
	230301366 - 230301380	intergenic	1	-
	172489402 - 172489416	intergenic	1	-
	26085001 - 26085015	intergenic	1	-
	13534895 - 13534909	intergenic	1	-
	9391062 - 9391076	intergenic	1	-
	5389439 - 5389453	intergenic	1	-
	247268230 - 247268244	intergenic	1	+
	186991257 - 186991271	intergenic	1	+
	29417463 - 29417477	intergenic	1	+
	214835896 - 214835910	intergenic	1	+
	36610115 - 36610129	intergenic	1	-
	211895732 - 211895746	intergenic	1	-
	201185097 - 201185111	intergenic	1	-
	110389309 - 110389324	start_codon, UTR	24	-
	215436450 - 215436464	intergenic	1	-
chr X	53333061 - 53333075	intergenic	1	+
	84524266 - 84524280	intergenic	1	+
	104432503 - 104432517	intergenic	1	+
	139483935 - 139483949	intergenic	1	+
	14503577 - 14503591	intergenic	1	-
	34447669 - 34447683	intergenic	1	-

Продолжение таблицы 7.

chr X	76444025 - 76444039	intergenic	1	-
	87251802 - 87251816	intergenic	1	-
	103638349 - 103638363	intergenic	1	-
	143666754 - 143666768	intergenic	1	-
	144912021 - 144912035	intergenic	1	-

Полный список результатов анализа всех участков мотива НАЕЕ представлен в Приложении Б.

3.2.1. Анализ межгенных регионов

Для межгенных регионов сделан анализ рамок считывания с помощью ручного скрипта, написанного на языке Python 3.12.7. Для каждого локуса анализировался окружающий геномный регион в диапазоне 300 нуклеотидов относительного изучаемого мотива. Полученные последовательности использовались для поиска потенциальных старт-кодонов (ATG, CTG, GTG, TTG), стоп-кодонов (TAA, TAG, TGA) во всех трех рамках считывания. В результате было выявлено 6238 потенциальных ORFs различной длины. Большинство найденных рамок считывания имело малую длину (1-5 нуклеотида) и, вероятно, являются нефункциональными участками.

Для дальнейшего анализа были отобраны ORFs длиной не менее 10 аминокислот, рассматриваемые как потенциальные кандидаты на sORFs, и содержащие мотив пептида НАЕЕГ. В результате остались 257 потенциальных источников эндогенного пептида НАЕЕГ с различными старт- и стоп-кодонами. В наиболее подходящих вариантах оставили кандидатов с длиной ORF 30-150 нуклеотидов и длиной пептида 10-50 а.о. (табл. 8).

Таблица 8. Наиболее интересные sORFs в межгенных областях исследуемого пептида HAEEG

Хромосома	Участок	Старт - кодон	Длина пептида (а.о)	Цепь ДНК (+/-)	Пептид
chr 1	9391062-9391076	TTG	15	-	LQRRFQQGSSHA EEG
chr 1	211895732-211895 746	TTG	16	+	LSGPHLCHHAEE GSEES
chr 1	211895732-211895 746	TTG	11	-	LCHHAEEEGSEES
chr 1	13534851-1353495 2	ATG	25	-	MVVVVSATHAEE GGLLEARRLLRLQ
chr 1	13534895-1353490 9	GTG	24	-	VVVVSATHAEEG GLLEARRLLRLQ
chr 2	86360714-8636072 8	CTG	30	+	LDFSSYFNSESIK AFSFSLHAEEGSV KAW
chr 2	27148854-2714886 8	TTG	27	-	LPPPP...HAEEGRR IVGK...
chr 3	165912521-165912 535	TTG	13	+	LGDHAEEGKVEQ N
chr 4	23977272-2397728 6	ATG	13	-	MNGFSAHAEEGG N
	23977272-2397728 6	GTG	16	-	VANMNGFSAHAEE EGGN
	182840218-182840 232	TTG	13	-	LGVRHAEEGKKSG
	182840218-182840 232	GTG	11	-	VRHAEEGKKSG
chr 6	1270921-1270935	ATG	24	-	MHAEEGRPHRSE MNNAILPQGILP

Продолжение таблицы 8.

chr 6	1270921-1270935	CTG	26	-	LGMHAE ^E GRPHR SEMNNAILPQGILP
chr 7	144101205-144101 219	TTG	13	-	LAQRIHAE ^E GRGN
chr 11	38612240-3861225 4	CTG	18	+	LSCGKTTGILLHA EEGFMN
chr 12	88654055-8865406 9	CTG	10	-	LLHAE ^E EGSPG
chr 13	44930516-4493053 0	ATG	16	-	MPMHAE ^E EGMPAS RCWCL
chr 17	8876001-8876015	TTG	11	+	LVRYHAE ^E EGPP
chr 18	65150022-6515003 6	CTG	10	+	LFHAE ^E EGHCG

Это участки ДНК между аннотированными генами, которые раньше называли «мусорной ДНК», известны сейчас способностью транскрибироваться в длинные некодирующие РНК (днкРНК). По разным оценкам, до 40% днкРНК могут транслироваться. Пока возможность трансляции удалось доказать только для немногих длинных некодирующих РНК, предшественников микроРНК и кольцевых РНК, но каждый год становятся известны все новые некодирующие РНК, участвующих в самых замысловатых молекулярных процессах. Кроме того, накапливается всё больше свидетельств того, что некоторые некодирующие РНК все-таки транслируются, правда, не в большие белки, а в короткие пептиды. Большинство из них, вероятно, нестабильны и быстро деградируют, но часть из них эволюционно консервативна и может выполнять регуляторные функции. К сожалению пока что современные методы не позволяют идентифицировать редкие или быстро разрушаемые пептиды, которые, тем не менее, могут быть физиологически значимы, к тому же, многие некодирующие РНК транслируются только в клетках определенного типа или при особых условиях. поэтому данные

результаты требуют дальнейшего более глубокого анализа с использованием инструментов, основанных на машинном обучении [54].

3.2.2. Анализ 5'UTR регионов и идентификация uORF

Для оценки возможности трансляции пептида НАЕЕ из нетранслируемых областей генов был проведен анализ 5'UTR-регионов. Из 197 мотивов, первоначально обнаруженных в UTR-участках, после фильтрации оставлены только 5'UTR-регионы с полноценной uORF (имеют старт- и стоп-кодон в одной рамке считывания). Для дальнейшего анализа были отобраны открытые рамки считывания с минимальной длиной 30 нуклеотидов (10 аминокислот), что соответствует общепринятым критериям для поиска коротких регуляторных пептидов. В результате были выявлены 29 uORF, принадлежащих 11 уникальным генам (табл. 9).

Идентифицированные uORF в регионе 5'UTR характеризуются:

- 1) Длиной пептидного продукта от 9 до 50 аминокислот;
- 2) Использованием как канонических (ATG), так и неканонических старт-кодонов (CTG, TTG, GTG);
- 3) Локализацией мотива НАЕЕГ во внутренней части или на концах предсказанного пептида.

Таблица 9. Уникальные гены с предсказанными uORF, содержащими мотив HAEFG, в 5'UTR

Хромосома	Координаты мотива	Ген	Функция белка	Цепь ДНК (+/-)	Старт-код	Длина uORF (нт)
chr 4	18289052-182890539	DCTD	Кодирует фермент дезоксицитидинмонофосфатдезаминазу (dCMP-дезаминазу)	+	CTG	150
chr 10	11335382-11335396	CELF2	Кодирует одноименный РНК-связывающий белок, который контролирует процессы сплайсинга, стабильность и трансляцию матричных РНК (мРНК) в клетках	-	TTG	150
					GTG	96
					CTG	87
	91281903-91281917	PCGF5	Кодирует регуляторный белок, который является ключевым компонентом белкового комплекса PRC1, отвечающего за подавление экспрессии определенных генов и поддержание эпигенетической памяти клеток	-	CTG	87
chr 11	17395868-17395882	ABCC8	Отвечает за выработку белка SUR1 (рецептора сульфонилмочевины 1-го типа)	+	GTG	153
	57645947-57645961	YPEL4	Кодирует белок, который играет важную роль в регуляции клеточного цикла, пролиферации и поддержании структуры клеточных мембран	+	GTG	135
					CTG	108
					GTG	93

Продолжение таблицы 9.

chr 11	11156098 5-1115609 99	LAYN	Кодирует белок лаилин (layilin) - трансмембранный рецептор, который связывает гиалуроновую кислоту и участвует во многих клеточных процессах	-	CTG	111
					CTG	105
					TTG	87
					ATG	57
					CTG	30
chr 12	49903867- 49903881	FAIM 2	Кодирует белок, который действует как мощный ингибитор апоптоза (запрограммированной клеточной гибели)	-	GTG	111
					ATG	144
chr 17	1464598-1 464612	MYO 1C	Кодирует белок миозин-1C - моторный белок, который связывает актиновый цитоскелет с клеточными мембранами и участвует в механическом ответе клеток	+	CTG	147
					CTG	141
					GTG	135
					GTG	129
chr 19	417192-41 7206	SHC2	Кодирует адаптерный белок, который передает сигналы от активированных рецепторов факторов роста внутрь клетки	+	CTG	123
					ATG	90
chr 20	353914-35 3928	NRSN 2	Кодирует нейрональный мембранный белок, который участвует в развитии нервной системы и регуляции внутриклеточных путей	-	GTG	147
					ATG	141
					CTG	99
					TTG	57
chr 22	29825108- 29825122	ASCC 2	Молекулярный адаптер в комплексе ASCC, участвующий в сплайсинге и репарации ДНК	+	ATG	108
					GTG	99
					GTG	78

В результате сопоставления 29 идентифицированных ORF из 5'UTR с ранее полученным списком из 37 белок-кодирующих генов видно, что для генов ABCC8 и ASCC2 мотив HAEE был обнаружен как в CDS, так и в 5'UTR участках. Это указывает на потенциальную двойную регуляцию продукции данного пептида: возможна его генерация как из основного белка (путем протеолиза), так и из короткого пептида, транслируемого с открытой рамки считывания. Наличие uORF также может свидетельствовать о сложных механизмах регуляции экспрессии этих генов на трансляционном уровне.

Интересно также отметить, что 8 из 29 идентифицированных ORF (27,6%) принадлежат транскриптам, аннотированным как “nonsense mediated decay” (NMD). Традиционно считается, что такие транскрипты не производят функциональных белков, так как быстро деградируют клеточными механизмами контроля качества. К найденной категории относятся гены ABCC8 и ASCC2, ранее выделенные как перспективные кандидаты на основании анализа CDS. Обнаружение потенциально кодирующих последовательностей в транскриптах, которые традиционно считаются «дефектными» и быстро деградирующими, не является артефактом. Напротив, современные данные указывают, что короткие открытые рамки считывания в 5'UTR таких транскриптов могут быть источником функциональных микропептидов [52,53].

Остальные найденные uORF в участках 9 генов (DCTD, CELF2, PCGF5, YPEL4, LAYN, FAIM2, MYO1C, NRSN2, SHC2) не были идентифицированы при первичном поиске по CDS. Обнаружение в их 5'UTR полноценных рамок считывания, содержащих мотив пептида HAEE, позволяет рассматривать их как новые, независимые источники эндогенного пептида, функционирующие, вероятно, через механизм uORF-опосредованной трансляции.

Таким образом в результате анализа наиболее перспективными кандидатами для дальнейшего экспериментального исследования признаны:

- 1) FAIM2 (нейропротекторный белок, экспрессирующийся в мозге);
- 2) NRSN2 (нейрон-специфичный белок синаптических везикул);
- 3) CELF2 (регулятор альтернативного сплайсинга в нейронах);

- 4) ABCC8 и ASCC2 (гены с двойным механизмом продукции HAEE через CDS и uORF, включая NMD-чувствительные транскрипты).

Отсутствие в литературе данных о функциональных открытых рамках считывания в этих генах подчеркивает новизну полученных результатов и открывает перспективы для дальнейших исследований в области микропептидов и их роли в нейродегенеративных процессах.

Для окончательного подтверждения трансляции идентифицированных открытых рамок считывания *in vivo* рекомендуется произвести анализ Ribo-seq данных из базы ENCODE (Encyclopedia of DNA Elements).

3.2.2.1. Контрольная валидация результатов с использованием BLASTN

Проверка корректности найденных локусов с использованием сервиса BLASTN базы данных NCBI (<https://blast.ncbi.nlm.nih.gov>) показала, что для всех 29 uORF наблюдается 100% идентичность соответствующих нуклеотидных последовательностей референсному геному человека.

Таким образом, валидация методом BLASTN полностью подтвердила корректность идентификации всех uORF и отсутствие артефактов в алгоритме их поиска.

ВЫВОДЫ

1. Проведен полногеномный поиск мотива HAEEG в геноме человека GRCh 38 в обеих цепях ДНК, выполнено аннотирование найденных локусов с использованием GTF-аннотации GENCODE v49 и проанализированы CDS-участки (37 находок), межгенные регионы (154 находок) и нетранслируемые участки (29 находок).
2. Сделано предположение, что пептид HAEEG не случайно встречается в геноме - он локализован в определенных локусах и может иметь функциональное значение для организма человека.
3. Присутствие мотива в CDS 37 генов предполагает, что HAEEG может образовываться из белков-предшественников. Наиболее вероятным источником эндогенного HAEEG являются гены *CACNA1B*, *BAIAP3* и *ABCC8*, которые экспрессируются в мозге и эндокринной системе и содержат мотив HAEEG во фрагменте.
4. Находки в межгенных регионах открывают возможность существования некодирующих РНК, транслируемых в короткие пептиды.
5. Предложены альтернативные uORF-опосредованные источники, включающие гены *FAIM2*, *NRSN2* и *CELF2*, в 5'UTR-участках которых обнаружены полноценные открытые рамки считывания. Особый интерес представили гены *ABCC8* и *ASCC2*, для которых обнаружен потенциально двойной механизм продукции HAEE (из основного белка и из uORF)

Полученные результаты свидетельствуют о широком распространении исследуемого мотива в геноме человека и указывают на необходимость дальнейшего анализа его локализации относительно аннотированных генов и найденных участков с использованием ML/AI инструментов (машинное обучение и искусственный интеллект (machine learning/artificial intelligence)).

СПИСОК ЛИТЕРАТУРЫ

1. Bermejo-Pareja F. и др. Incidence and subtypes of dementia in three elderly populations of central Spain // J Neurol Sci. 2008. V. 264, № 1–2. P. 63–72.
2. Mendez M. F. Early-onset Alzheimer Disease and Its Variants // Continuum (Minneap Minn). Minneapolis, Minn., 2019. V. 25, № 1. P. 34–51.
3. Dementia [Электронный ресурс]. URL: <https://www.who.int/news-room/fact-sheets/detail/dementia> (дата обращения: 06.03.2026).
4. Knopman D. S. и др. Alzheimer disease // Nat Rev Dis Primers. 2021. V. 7, № 1. P. 33.
5. Sperling R. A. и др. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease // Alzheimers Dement. 2011. V. 7, № 3. P. 280–292.
6. DeTure M. A., Dickson D. W. The neuropathological diagnosis of Alzheimer's disease // Mol Neurodegener. 2019. V. 14, № 1. P. 32.
7. Bloom G. S. Amyloid- β and tau: the trigger and bullet in Alzheimer disease pathogenesis // JAMA Neurol. 2014. V. 71, № 4. P. 505–508.
8. Selkoe D. J., Hardy J. The amyloid hypothesis of Alzheimer's disease at 25 years // EMBO Mol Med. 2016. V. 8, № 6. P. 595–608.
9. Butterfield D. A., Halliwell B. Oxidative stress, dysfunctional glucose metabolism and Alzheimer disease // Nat Rev Neurosci. 2019. V. 20, № 3. P. 148–160.
10. Braak H., Braak E. Neuropathological staging of Alzheimer-related changes // Acta Neuropathol. 1991. V. 82, № 4. P. 239–259.
11. Karran E., De Strooper B. The amyloid hypothesis in Alzheimer disease: new insights from new therapeutics // Nat Rev Drug Discov. 2022. V. 21, № 4. P. 306–318.
12. Scheltens P. и др. Alzheimer's disease // Lancet. London, England, 2021. V. 397, № 10284. P. 1577–1590.

13. Вестник РГМУ [Электронный ресурс]. URL: <https://vestnik.rsmu.press/archive/2026/1/5/media?lang=ru> (дата обращения: 06.03.2026).
14. Walker L. C. A β Plaques // Free Neuropathol. 2020. V. 1. P. 31.
15. Haass C., Selkoe D. J. Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid beta-peptide // Nat Rev Mol Cell Biol. 2007. V. 8, № 2. P. 101–112.
16. Benilova I., Karran E., De Strooper B. The toxic A β oligomer and Alzheimer's disease: an emperor in need of clothes // Nat Neurosci. 2012. V. 15, № 3. P. 349–357.
17. Heneka M. T. и др. Neuroinflammation in Alzheimer's disease // Lancet Neurol. 2015. V. 14, № 4. P. 388–405.
18. Berridge M. J. Calcium signalling and Alzheimer's disease // Neurochem Res. 2011. V. 36, № 7. P. 1149–1156.
19. Reddy P. H. и др. MicroRNAs, Aging, Cellular Senescence, and Alzheimer's Disease // Prog Mol Biol Transl Sci. 2017. V. 146. P. 127–171.
20. Pagani L., Eckert A. Amyloid-Beta interaction with mitochondria // Int J Alzheimers Dis. 2011. V. 2011. P. 925050.
21. Hardy J., Selkoe D. J. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics // Science. New York, N.Y., 2002. V. 297, № 5580. P. 353–356.
22. Zheng H., Koo E. H. Biology and pathophysiology of the amyloid precursor protein // Mol Neurodegener. 2011. V. 6, № 1. P. 27.
23. Not just amyloid: physiological functions of the amyloid precursor protein family | Nature Reviews Neuroscience [Электронный ресурс]. URL: <https://www.nature.com/articles/nrn.2017.29> (дата обращения: 06.03.2026).
24. Amyloid beta: structure, biology and structure-based therapeutic development | Acta Pharmacologica Sinica [Электронный ресурс]. URL: <https://www.nature.com/articles/aps201728> (дата обращения: 06.03.2026).
25. Postina R. Activation of α -secretase cleavage // J Neurochem. 2012. V. 120 Suppl 1. P. 46–54.

26. Marlow L. и др. Beta-secretase processing of the Alzheimer's amyloid protein precursor (APP) // J Mol Neurosci. 2003. V. 20, № 3. P. 233–239.
27. Vassar R. и др. Beta-secretase cleavage of Alzheimer's amyloid precursor protein by the transmembrane aspartic protease BACE // Science. New York, N.Y., 1999. V. 286, № 5440. P. 735–741.
28. Hardy J. A., Higgins G. A. Alzheimer's disease: the amyloid cascade hypothesis // Science. New York, N.Y., 1992. V. 256, № 5054. P. 184–185.
29. Tjernberg L. O. и др. Arrest of beta-amyloid fibril formation by a pentapeptide ligand // J Biol Chem. 1996. V. 271, № 15. P. 8545–8548.
30. Anashkina A. A. и др. Proteins with the HAEE tetrapeptide motif as potential targets for beta-amyloid // Research Results in Pharmacology. 2025. V. 11, № 4. P. 212–221.
31. Zolotarev Y. A. и др. Pharmacokinetics and Molecular Modeling Indicate nAChR α 4-Derived Peptide HAEE Goes through the Blood-Brain Barrier // Biomolecules. 2021. V. 11, № 6. P. 909.
32. Tsvetkov P. O. и др. Peripherally Applied Synthetic Tetrapeptides HAEE and RADD Slow Down the Development of Cerebral β -Amyloidosis in A β PP/PS1 Transgenic Mice // J Alzheimers Dis. 2015. V. 46, № 4. P. 849–853.
33. Mukhina K. A. и др. The Tetrapeptide HAEE Promotes Amyloid-Beta Clearance from the Brain // Int J Mol Sci. 2025. V. 26, № 23. P. 11591.
34. Craik D. J. и др. The future of peptide-based drugs // Chem Biol Drug Des. 2013. V. 81, № 1. P. 136–147.
35. Kechko O. I. и др. Therapeutic Potential of HAEEPGP Peptide Against β -Amyloid Induced Neuropathology // Mol Neurobiol. 2025. V. 63, № 1. P. 36.
36. Mitkevich V. A. и др. Zn-dependent β -amyloid Aggregation and its Reversal by the Tetrapeptide HAEE // Aging Dis. 2023. V. 14, № 2. P. 309–318.
37. Beyond Alzheimer's story: how an engineered molecule gained an endogenous essence | bioRxiv [Электронный ресурс]. URL: <https://www.biorxiv.org/content/10.1101/2025.09.22.677901v1.abstract> (дата обращения: 06.03.2026).

38. Checco J. W. Identifying and Measuring Endogenous Peptides through Peptidomics // ACS Chem Neurosci. 2023. V. 14, № 20. P. 3728–3731.
39. Kay R. G. и др. Peptidomic analysis of endogenous plasma peptides from patients with pancreatic neuroendocrine tumours // Rapid Commun Mass Spectrom. 2018. V. 32, № 16. P. 1414–1424.
40. Brennan A., Church T. R., Margolis S. S. Proteasome-derived peptides: separating the trash from the recycling // Trends Biochem Sci. 2025. V. 50, № 9. P. 779–794.
41. Saric T., Graef C. I., Goldberg A. L. Pathway for degradation of peptides generated by proteasomes: a key role for thimet oligopeptidase and other metallopeptidases // J Biol Chem. 2004. V. 279, № 45. P. 46723–46732.
42. Network analysis of the proteome and peptidome sheds light on human milk as a biological system | Scientific Reports [Электронный ресурс]. URL: <https://www.nature.com/articles/s41598-024-58127-2> (дата обращения: 06.03.2026).
43. Romanova E. V., Dowd S. E., Sweedler J. V. Quantitation of endogenous peptides using mass spectrometry based methods // Curr Opin Chem Biol. 2013. V. 17, № 5. P. 801–808.
44. Apostolopoulos V. и др. A Global Review on Short Peptides: Frontiers and Perspectives // Molecules. Basel, Switzerland, 2021. V. 26, № 2. P. 430.
45. Paes W. и др. Elucidation of the Signatures of Proteasome-Catalyzed Peptide Splicing // Front Immunol. 2020. V. 11. P. 563800.
46. Li Y. и др. Simple molecules make difference: short peptides play a novel role in slowing senescence // NPJ Aging. 2025. V. 11, № 1. P. 104.
47. Emerging evidence for functional peptides encoded by short open reading frames | Nature Reviews Genetics [Электронный ресурс]. URL: <https://www.nature.com/articles/nrg3520> (дата обращения: 06.03.2026).
48. Dong X. и др. Small Open Reading Frame-Encoded Micro-Peptides: An Emerging Protein World // Int J Mol Sci. 2023. V. 24, № 13. P. 10562.
49. Su H., Katz S. G., Slavoff S. A. Nested small open reading frames are translated from alternative transcripts // bioRxiv. 2025. P. 2024.10.22.619581.

50. Kozin S. A. и др. PepString server as a tool to search for short amino acid subsequences: identification of potential amyloid-beta targets // *Acta Naturae*. 2025. V. 17, № 3. P. 67–76.
51. Kumar D., Mains R. E., Eipper B. A. 60 YEARS OF POMC // *J Mol Endocrinol*. 2016. V. 56, № 4. P. T63–T76.
52. Lacerda R. и др. The uORF–NMD axis and uPeptides: a new frontier in cancer therapy // *Front. RNA Res. Frontiers*, 2026. V. 4.
53. Patro A. K. и др. Nonsense-mediated mRNA decay: Physiological significance, mechanistic insights and future implications // *Pathol Res Pract*. 2024. V. 264. P. 155677.
54. Li L.-J. и др. Translation of noncoding RNAs: Focus on lncRNAs, pri-miRNAs, and circRNAs // *Experimental Cell Research*. 2017. V. 361, № 1. T. 1–8.

СПИСОК СОКРАЩЕНИЙ

A β - бета-амилоидный пептид (beta-amyloid peptide)

APP - амилоидный белок-предшественник (amyloid precursor protein)

BLASTN - инструмент поиска гомологичных нуклеотидных последовательностей (Basic Local Alignment Search Tool for nucleotides)

CDS - кодирующая последовательность (coding sequence)

FASTA - текстовый формат для хранения биологических последовательностей (FASTA format)

GENCODE - совместный проект аннотации генов человека и мыши (GENCODE gene annotation)

GRCh38 - референсная сборка генома человека (Genome Reference Consortium human build 38)

GTEx - проект по картированию экспрессии генов в тканях человека (Genotype-Tissue Expression project)

GTF - формат файла для описания аннотации генов (Gene Transfer Format)

IUPAC - Международный союз теоретической и прикладной химии (International Union of Pure and Applied Chemistry)

lncRNA - длинная некодирующая РНК (long non-coding RNA)

NCBI - Национальный центр биотехнологической информации (National Center for Biotechnology Information)

ORF - открытая рамка считывания (open reading frame)

Ribo-seq - метод секвенирования рибосом-защищенных фрагментов РНК (ribosome profiling sequencing)

sORF - короткая открытая рамка считывания (short open reading frame)

UniProt - универсальная база данных белков (Universal Protein Resource)

uORF - короткая открытая рамка считывания выше основного старт-кодона гена (upstream open reading frame)

UTR - нетранслируемая область (untranslated region)

Приложение А

Результаты BLASTN-валидации

Ген	Последовательность мотива	Длина и идентичность	e-value	Координаты совпадения
GPS1	CACGCGGAAGAA GGC	15/15 (100%)	6,4	chr17:82054587-82054602
ASCC2	CATGCGGAGGAA GGT	15/15 (100%)	6,4	chr22:29825107-29825122
BAIAP3	CATGCTGAAGAA GGG	15/15 (100%)	6,4	chr16:1338943-1338958
NUMA1	GCCTTCTTCAGCA TG	15/15 (100%)	6,4	chr11:72013174-72013189
SMPD3	CCCCTCCTCTGCA TG	15/15 (100%)	6,4	chr16:68361623-68361638
DMPK	CCCCTCTTCAGCA TG	15/15 (100%)	6,4	chr19:45779442-45779457
ZBTB48	CATGCAGAAGAA GGA	15/15 (100%)	6,4	chr1:6587227,6587242
FMO3	CATGCAGAGGAG GGC	15/15 (100%)	6,4	chr1:171103787,171103802
DLL1	CACGCGGAAGAA GGT	15/15 (100%)	6,4	chr6:170289676,170289691
ABCC8	CATGCGGAAGAA GGC	15/15 (100%)	6,4	chr11:17395867,17395882
MAP3K5	TCCTTCCTCAGCA TG	15/15 (100%)	6,4	chr6:136592257,136592272
CRACR2 A	GCCTTCCTCAGCA TG	15/15 (100%)	6,4	chr12:3696835,3696850
CACNA1 B	CACGCAGAGGAG GGA	15/15 (100%)	6,4	chr9:137957641,137957656

Продолжение приложения А.

SLC16A4	TCCCTCTTCAGCA TG	15/15 (100%)	6,4	chr1:110389309,11038 9324
GAS8	CATGCGGAAGAA GGA	Неподтвержденный кандидат		chr16:90037245,9003 7260
DGKQ	GCCTTCCTCAGCA TG	15/15 (100%)	6,4	chr4:960701,960716
CMTM7	ACCTTCCTCTGCA TG	15/15 (100%)	6,4	chr3:32452409,32452 424
CATIP	CACGCGGAGGAG GGC	15/15 (100%)	6,4	chr2:218364699,2183 64714
LRRC23	CATGCTGAAGAA GGT	15/15 (100%)	6,4	chr12:6909897,69099 12
SEC14L6	ACCTTCCTCAGCA TG	Неподтвержденный кандидат		chr22:30533993,3053 4008
ZNF746	CATGCTGAAGAG GGA	15/15 (100%)	6,4	chr7:149476965,1494 76980
FYCO1	TCCTTCCTCAGCA TG	15/15 (100%)	6,4	chr3:45964371,45964 386
CNGA3	CACGCTGAAGAA GGT	15/15 (100%)	6,4	chr2:98396594,98396 609
RETREG 1	TCCCTCCTCGGCG TG	15/15 (100%)	6,4	chr5:16616932,16616 947
GPATCH 8	CCCTTCCTCCGCA TG	15/15 (100%)	6,4	chr17:44401098,4440 1113
G0S2	GCCCTCTTCGGCG TG	15/15 (100%)	6,4	chr1:209675783,2096 75798
PRRC2B	CCCCTCCTCGGCA TG	15/15 (100%)	6,4	chr9:131446509,1314 46524
KMT2B	CCCCTCCTCGGCG TG	15/15 (100%)	6,4	chr19:35733260,3573 3275
ABTB2	CATGCTGAGGAG GGG	15/15 (100%)	6,4	chr11:34164762,34164 777

Продолжение приложения А.

GOLT1A	CATGCCGAGGAG GGG	15/15 (100%)	6,4	chr1:204201667,204201682
OR5M10	CATGCAGAAGAG GGT	15/15 (100%)	6,4	chr11:56576947,56576962
OR5M1	CATGCAGAAGAG GGT	15/15 (100%)	6,4	chr11:56612728,56612743
AGBL1	ACCTTCTTCTGCA TG	15/15 (100%)	6,4	chr15:86988066,86988081
CHST10	GCCTTCCTCTGCA TG	15/15 (100%)	6,4	chr2:100417015,100417030
ASPSCR 1	CCCCTCCTCAGCA TG	Неподтвержденный кандидат		chr17:81976916,81976931
SYCE1	CACGCAGAGGAA GGC	Неподтвержденный кандидат		chr10:133568407,133568422

Приложение Б

Анализ локализации мотива НАЕЕ в некодирующих регионах

Хромосома	Участок	Участок генома	Количество изоформ мотива	Цепь ДНК (+/-)
chr 1	51300371 - 51300386	UTR	1	+
	171103787 - 171103802	UTR	1	+
	230301366 - 230301380	intergenic	1	-
	172489402 - 172489416	intergenic	1	-
	26085001 - 26085015	intergenic	1	-
	13534895 - 13534909	intergenic	1	-
	9391062 - 9391076	intergenic	1	-
	5389439 - 5389453	intergenic	1	-
	247268230 - 247268244	intergenic	1	+
	186991257 - 186991271	intergenic	1	+
	29417463 - 29417477	intergenic	1	+
	214835896 - 214835910	intergenic	1	+
	36610115 - 36610129	intergenic	1	-
	211895732 - 211895746	intergenic	1	-
	201185097 - 201185111	intergenic	1	-
	110389309 - 110389324	start_codon, UTR	24	-
	215436450 - 215436464	intergenic	1	-
chr 2	100417015 - 100417030	start_codon, UTR	11	-
	236724217 - 236724231	intergenic	1	-
	67855021 - 67855035	intergenic	1	+

Продолжение приложения Б.

chr 2	100417015 - 100417030	start_codon, UTR	11	-
	236724217 - 236724231	intergenic	1	-
	67855021 - 67855035	intergenic	1	+
	178188682 - 178188696	intergenic	1	-
	86360714 - 86360728	intergenic	1	+
	16180711 - 16180725	intergenic	1	-
	133734447 - 133734461	intergenic	1	+
	157658183 - 157658197	intergenic	1	-
	112261286 - 112261300	intergenic	1	-
	27148854 - 27148868	intergenic	1	-
	147437226 - 147437240	intergenic	1	-
chr 3	32452409 - 32452424	UTR	1	-
	137977439 - 137977453	intergenic	1	-
	129065646 - 129065660	intergenic	1	-
	123031896 - 123031910	intergenic	1	-
	67976330 - 67976344	intergenic	1	-
	129058189 - 129058203	intergenic	1	+
	78084431 - 78084445	intergenic	1	+
	33232320 - 33232334	intergenic	1	-
	165912521 - 165912535	intergenic	1	+
chr 4	182890524 - 182890539	UTR	26	+
	182840218 - 182840232	intergenic	1	-
	115500464 - 115500478	intergenic	1	-
	179698377 - 179698391	intergenic	1	+
	117312272 - 117312286	intergenic	1	+

Продолжение приложения Б.

chr 4	28115451 - 28115465	intergenic	1	+
	10263188 - 10263202	intergenic	1	+
	67247777 - 67247791	intergenic	1	-
	189416379 - 189416393	intergenic	1	+
	182890524 - 182890539	UTR	26	+
	182840218 - 182840232	intergenic	1	-
	115500464 - 115500478	intergenic	1	-
	179698377 - 179698391	intergenic	1	+
	117312272 - 117312286	intergenic	1	+
	28115451 - 28115465	intergenic	1	+
	10263188 - 10263202	intergenic	1	+
	67247777 - 67247791	intergenic	1	-
	189416379 - 189416393	intergenic	1	+
	23977272 - 23977286	intergenic	1	-
	38563684 - 38563698	intergenic	1	-
	38783039 - 38783053	intergenic	1	-
	152593941 - 152593955	intergenic	1	-
	69544346 - 69544360	intergenic	1	-
	120011449 - 120011463	intergenic	1	-
chr 5	132867754 - 132867769	UTR	1	-
	166668188 - 166668202	intergenic	1	-
	25488523 - 25488537	intergenic	1	+
	5903103 - 5903117	intergenic	1	+
	173969780 - 173969794	intergenic	1	+
	153256737 - 153256751	intergenic	1	+

Продолжение приложения Б.

chr 5	133900567 - 133900581	intergenic	1	+
	2813027 - 2813041	intergenic	1	-
	180230297 - 180230311	intergenic	1	+
	133770927 - 133770941	intergenic	1	-
	23690444 - 23690458	intergenic	1	-
chr 6	159437015 - 159437029	intergenic	1	-
	122941347 - 122941361	intergenic	1	-
	82468589 - 82468603	intergenic	1	-
	135146499 - 135146513	intergenic	1	+
	14111861 - 14111875	intergenic	1	+
	80106491 - 80106505	intergenic	1	-
	141501281 - 141501295	intergenic	1	-
	62647673 - 62647687	intergenic	1	-
	1270921 - 1270935	intergenic	1	-
	141456340 - 141456354	intergenic	1	-
chr 7	151189219 - 151189233	intergenic	1	-
	120871293 - 120871307	intergenic	1	-
	119417886 - 119417900	intergenic	1	-
	55743791 - 55743805	intergenic	1	-
	146093896 - 146093910	intergenic	1	+
	67949018 - 67949032	intergenic	1	+
	142670344 - 142670358	intergenic	1	-
	144101205 - 144101219	intergenic	1	-
chr 8	141672631 - 141672645	intergenic	1	-
	53691068 - 53691082	intergenic	1	+

Продолжение приложения Б.

chr 8	5675373 - 5675387	intergenic	1	+
	20749155 - 20749169	intergenic	1	+
	30771931 - 30771945	intergenic	1	+
	126331697 - 126331711	intergenic	1	+
	98339745 - 98339759	intergenic	1	-
	1965741 - 1965755	intergenic	1	-
	14075773 - 14075787	intergenic	1	-
	133407920 - 133407934	intergenic	1	-
chr 9	134236220 - 134236234	intergenic	1	-
	130543308 - 130543322	intergenic	1	-
	16068625 - 16068639	intergenic	1	-
	131815146 - 131815160	intergenic	1	-
	122670218 - 122670232	intergenic	1	-
chr 10	11335381 - 11335396	UTR	3	-
	91281902 - 91281917	UTR	8	-
	1860720 - 1860734	intergenic	1	+
	774450 - 774464	intergenic	1	+
	89173262 - 89173276	intergenic	1	+
	43712070 - 43712084	intergenic	1	+
	80818776 - 80818790	intergenic	1	+
	113500051 - 113500065	intergenic	1	+
	81287086 - 81287100	intergenic	1	-
	56656194 - 56656208	intergenic	1	-
	99712875 - 99712890	UTR	3	-
	128884789 - 128884803	intergenic	1	-

Продолжение приложения Б.

chr 11	17395867 - 17395882	UTR	27	+
	57645946 - 57645961	UTR	5	+
	93987698 - 93987712	intergenic	1	-
	11222158 - 11222172	intergenic	1	+
	30359227 - 30359241	intergenic	1	+
	38612240 - 38612254	intergenic	1	+
	104701274 - 104701288	intergenic	1	+
	111560984 - 111560999	UTR	4	-
chr 12	6909897 - 6909912	UTR	2	+
	100420544 - 100420559	UTR	2	+
	49903866 - 49903881	UTR	6	-
	86919012 - 86919026	intergenic	1	+
	88858518 - 88858532	intergenic	1	-
	60746332 - 60746346	intergenic	1	+
	17922253 - 17922267	intergenic	1	+
	104208090 - 104208104	intergenic	1	+
chr 13	44930516 - 44930530	intergenic	1	-
	56898204 - 56898218	intergenic	1	+
	30430394 - 30430408	intergenic	1	+
	99131513 - 99131527	intergenic	1	+
	76357095 - 76357109	intergenic	1	-
	89913354 - 89913368	intergenic	1	-
chr 14	98279172 - 98279186	intergenic	1	-
	81678102 - 81678116	intergenic	1	-
	101107698 - 101107712	intergenic	1	+

Продолжение приложения Б.

chr 14	49294759 - 49294773	intergenic	1	-
	23105737 - 23105751	intergenic	1	+
	46777890 - 46777904	intergenic	1	-
chr 15	55194917 - 55194932	UTR	1	+
	26545564 - 26545579	UTR	5	-
	84969373 - 84969387	intergenic	1	+
chr 16	90037245 - 90037260	UTR	1	+
	68361623 - 68361638	UTR	1	-
	1464597 - 1464612	UTR	5	+
	82054587 - 82054602	UTR	4	+
	44401098 - 44401113	UTR	1	-
	79405059 - 79405073	intergenic	1	-
chr 17	81493095 - 81493109	intergenic	1	-
	44401099 - 44401113	UTR	1	-
	56653829 - 56653843	intergenic	1	+
	12263663 - 12263677	intergenic	1	+
	8876001 - 8876015	intergenic	1	+
	1464598 - 1464612	UTR	5	+
	82054588 - 82054602	UTR	4	+
	75135482 - 75135496	UTR	6	-
	34435408 - 34435422	intergenic	1	-
chr 18	78644237 - 78644251	intergenic	1	-
	74367427 - 74367441	intergenic	1	-
	35410242 - 35410256	intergenic	1	+
	29393618 - 29393632	intergenic	1	+

Продолжение приложения Б.

chr 18	65150022 - 65150036	intergenic	1	+
chr 19	417191 - 417206	UTR	14	+
	33780888 - 33780902	intergenic	1	-
	30094365 - 30094379	intergenic	1	-
	29694494 - 29694508	intergenic	1	-
	35733260 - 35733275	UTR	2	-
	45779442 - 45779457	UTR	2	-
chr 20	353913 - 353928	UTR	27	-
	55137896 - 55137910	intergenic	1	-
	64046022 - 64046036	intergenic	1	-
chr 21	28243542 - 28243556	intergenic	1	-
chr 22	29825107 - 29825122	UTR	2	+
	20174561 - 20174575	intergenic	1	+
	30533993 - 30534008	start_codon, UTR	6	-
	36186432 - 36186446	intergenic	1	+
	39563066 - 39563080	intergenic	1	+
	50253047 - 50253062	UTR	10	-
chr X	53333061 - 53333075	intergenic	1	+
	84524266 - 84524280	intergenic	1	+
	104432503 - 104432517	intergenic	1	+
	139483935 - 139483949	intergenic	1	+
	14503577 - 14503591	intergenic	1	-
	34447669 - 34447683	intergenic	1	-
	76444025 - 76444039	intergenic	1	-
	87251802 - 87251816	intergenic	1	-

Продолжение приложения Б.

chr X	103638349 - 103638363	intergenic	1	-
	143666754 - 143666768	intergenic	1	-
	144912021 - 144912035	intergenic	1	-