https://doi.org/10.21518/ms2026-095

Оригинальная статья / Original article

Инструментальная биоинформатика в персонифицированной медицине: от алгоритмов выравнивания к самостоятельному врачебному анализу геномных вариаций в среде R

А.А. Корнеенков[1]✉, https://orcid.org/0000-0001-5870-8042, korneyenkov@gmail.com
Ю.К. Янов[2,3], https://orcid.org/0000-0001-9195-128X, 9153864@mail.ru
Е.Э. Вяземская[1], https://orcid.org/0000-0002-4141-2226, vyazemskaya.elena@gmail.com
А.Ю. Медведева[1], https://orcid.org/0009-0002-6921-5299, a.medvedeva@niilor.ru

[1] Санкт-Петербургский научно-исследовательский институт уха, горла, носа и речи; 190013, Россия, Санкт-Петербург, ул. Бронницкая, д. 9

[2] Военно-медицинская академия имени С.М. Кирова; 194044, Россия, Санкт-Петербург, ул. Академика Лебедева, д. 6

[3] Северо-Западный государственный медицинский университет имени И.И. Мечникова; 191015, Россия, Санкт-Петербург, ул. Кирочная, д. 41

Резюме

Введение. Современная персонифицированная медицина требует от врача навыков самостоятельной интерпретации генетических вариантов. Инструментальная биоинформатика в среде R предоставляет специалисту мощный аппарат для верификации подозрительных находок. Оценка эволюционной консервативности аминокислотных позиций с помощью алгоритмов выравнивания является критическим этапом в определении клинической значимости миссенс-вариантов (согласно критериям ACMG).

Цель. Продемонстрировать алгоритм самостоятельного биоинформатического анализа в среде R для оценки патогенности мутации V37I в гене *GJB2*, ассоциированной с наследственной тугоухостью.

Материалы и методы. В работе использованы пакеты Bioconductor (Biostrings, pwalign, msa) и среда R. Материалом послужили девять полноразмерных ортологичных последовательностей белка коннексина 26 (CXB2), полученных из базы данных UniProt, охватывающих таксономические группы приматов, грызунов и парнокопытных. Реализован двухэтапный анализ: парное выравнивание для идентификации замены у пациента и множественное выравнивание (multiple sequence alignment, MSA) для расчета индекса консервативности локуса.

Результаты. На примере анализа миссенс-варианта V37I (ген *GJB2*) продемонстрирована работоспособность двухэтапного алгоритма выравнивания в среде R. С помощью парного выравнивания (пакет pwalign) успешно идентифицирована замена нуклеотида, приводящая к аминокислотному сдвигу (PID = 93,75%). MSA последовательностей 9 видов млекопитающих позволило наглядно визуализировать абсолютную инвариантность 37-й позиции белка в дикой природе. Расчет индекса консервативности (0,9) подтвердил возможность автоматизированного получения данных для оценки варианта по критерию PP3 (ACMG). Предложенный программный подход позволяет врачу-исследователю самостоятельно переходить от «сырых» данных UniProt к экспертной визуализации без использования громоздких вычислительных комплексов.

Заключение. Самостоятельное использование врачом инструментов биоинформатики в среде R позволяет перейти от пассивного изучения лабораторных отчетов к активному анализу геномных данных. Продемонстрированный подход обеспечивает высокую доказательность клинических выводов и является важным элементом системы поддержки принятия врачебных решений в рамках персонифицированной медицины.

Ключевые слова: биоинформатика в медицине, Bioconductor, ген *GJB2*, коннексин 26, множественное выравнивание (MSA), консервативность аминокислот, критерии ACMG

Для цитирования: Корнеенков АА, Янов ЮК, Вяземская ЕЭ, Медведева АЮ. Инструментальная биоинформатика в персонифицированной медицине: от алгоритмов выравнивания к самостоятельному врачебному анализу геномных вариаций в среде R. *Медицинский совет.* 2026;20(6):161–168. https://doi.org/10.21518/ms2026-095.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Bioinformatics tools for personalized medicine: From alignment algorithms to physician-driven analysis of genomic variations in the R environment

Aleksei A. Korneenkov[1]✉, https://orcid.org/0000-0001-5870-8042, korneyenkov@gmail.com
Yuri K. Yanov[2,3], https://orcid.org/0000-0001-9195-128X, 9153864@mail.ru
Elena E. Vyazemskaya[1], https://orcid.org/0000-0002-4141-2226, vyazemskaya.elena@gmail.com
Anna Yu. Medvedeva[1], https://orcid.org/0009-0002-6921-5299, a.medvedeva@niilor.ru

[1] Saint Petersburg Research Institute of Ear, Throat, Nose and Speech; 9, Bronnitskaya St., St Petersburg, 190013, Russia
[2] Kirov Military Medical Academy; 6, Akademik Lebedev St., St Petersburg, 194044, Russia
[3] North-Western State Medical University named after I.I. Mechnikov; 41, Kirochnaya St., St Petersburg, 191015, Russia

Abstract
Introduction. Modern personalized medicine requires physicians to possess skills for the independent interpretation of genetic variants. Bioinformatics tools within the R environment provide a powerful framework for the verification of suspicious findings. The assessment of the evolutionary conservation of amino acid positions using alignment algorithms is a critical step in determining the clinical significance of missense variants (according to ACMG criteria).
Aim. To demonstrate a standalone bioinformatics analysis algorithm in the R environment for assessing the pathogenicity of the V37I mutation in the *GJB2* gene, associated with hereditary hearing loss.
Materials and methods. This study utilized Bioconductor packages (Biostrings, pwalign, msa) within the R environment. The material comprised nine full-length orthologous sequences of the connexin 26 protein (CXB2), obtained from the UniProt database, covering the taxonomic groups of primates, rodents, and even-toed ungulates. A two-step analysis was implemented: pairwise alignment to identify the substitution in the patient and multiple sequence alignment (MSA) to calculate the conservation index of the locus.
Results. Using the analysis of the missense variant V37I (*GJB2* gene) as an example, the functionality of the two-step alignment algorithm in the R environment was demonstrated. Pairwise alignment (pwalign package) successfully identified the nucleotide substitution leading to the amino acid change (PID = 93.75%). MSA of sequences from 9 mammalian species allowed for a clear visualization of the absolute invariance of the protein's 37th position in the wild-type state. Calculation of the conservation index (0.9) confirmed the feasibility of automated data acquisition for variant evaluation according to the PP3 criterion (ACMG). The proposed computational approach enables a clinician-researcher to independently transition from "raw" UniProt data to expert-level visualization without the need for cumbersome computational systems.
Conclusion. The independent use of bioinformatics tools in the R environment by physicians facilitates a shift from the passive interpretation of laboratory reports to the active analysis of genomic data. The demonstrated approach provides a high level of evidence for clinical conclusions and represents an important element of the clinical decision support system within the framework of personalized medicine.

Keywords: bioinformatics in medicine, Bioconductor, *GJB2* gene, connexin 26, multiple sequence alignment, amino acid conservation, ACMG criteria

For citation: Korneenkov AA, Yanov YuK, Vyazemskaya EE, Medvedeva AYu. Bioinformatics tools for personalized medicine: From alignment algorithms to physician-driven analysis of genomic variations in the R environment. *Meditsinskiy Sovet*. 2026;20(6):161–168. (In Russ.) https://doi.org/10.21518/ms2026-095.

Conflict of interest: the authors declare no conflict of interest.

ВВЕДЕНИЕ

Современная генетическая диагностика представляет собой сложный технологический конвейер биоинформатических методов (англ. *bioinformatics pipeline*). Он состоит из последовательных этапов: от загрузки «сырых» данных секвенирования до клинической интерпретации выявленных отклонений [1]. Внутри этого конвейера скрыт критически важный этап, определяющий успех всей персонифицированной медицины, но часто остающийся для врача «черным ящиком». Это – выравнивание последовательностей (англ. *sequence alignment*): математическое сопоставление двух или более цепочек нуклеотидов ДНК или аминокислот белков [2]. Именно на этом этапе происходит переход от анонимного кода к выявлению индивидуального генетического профиля, позволяющего врачу увидеть конкретные «опечатки», ставшие причиной патологии у данного пациента.

Выравнивание последовательностей может решать разные задачи: картирование, детекция отклонений, оценка вариабельности остатков в сопоставляемых позициях. Понимание принципов его работы позволяет врачу оценить достоверность найденных генетических вариантов, а не слепо доверять заключению лаборатории [3]. Выравнивание – это «сердце» биоинформатики и, пожалуй, самый высокий интеллектуальный барьер для клинициста. Раньше компетенции в этой области относились практически исключительно к молекулярной биологии, биоинформатике, генетике и другим узким специальностям. Однако сегодня информационные границы стали условными. В данной статье мы покажем, что, несмотря на математическую сложность, этот процесс может быть прозрачным и управляемым инструментом в руках врача, позволяющим не просто «получить результат», а попытаться осознать молекулярную причину болезни [4].

Цель исследования – продемонстрировать алгоритм выравнивания последовательностей в программной среде R как доступный инструмент самостоятельного биоинформатического анализа в практике врача-клинициста (на примере оценки консервативности мутации V37I в гене *GJB2*).

Для достижения этой цели определены следующие задачи:

- описать алгоритмы выравнивания последовательностей (парного, множественного), оценки гомологии и степени сходства участков (консервативности), схемы возможной дивергенции (филогения);
- сформировать программный код в среде R, иллюстрирующий эти процессы, на примере анализа белка коннексина 26 (*GJB2*) – ключевого маркера наследственной тугоухости [5–8].

Весь программный код, иллюстрирующий результаты, выполнен на языке R с использованием программных пакетов открытого исходного кода проекта Bioconductor. Несмотря на то что специализированные консольные утилиты (BWA, GATK) обладают более высокой производительностью для полногеномных данных, среда R предоставляет врачу-исследователю уникальные возможности для самостоятельной интерактивной визуализации, статистического контроля и воспроизводимости анализа на уровне отдельных генов и белков [1].

МАТЕРИАЛЫ И МЕТОДЫ

В основе биоинформатического анализа генетических вариантов лежит процедура выравнивания последовательностей. Выравниванием (англ. *alignment*) последовательностей называют определение взаимного соответствия оснований или остатков в двух или нескольких последовательностях, при котором сохраняется исходный порядок остатков в последовательностях. Две последовательности можно «выровнять», отобразив их гомологичные (схожие) нуклеотидные или аминокислотные остатки друг под другом в две строки, представляя их с помощью букв алфавита (A, C, T, G, U и т. д.). Выравнивание не должно изменять «смысл» последовательностей, поэтому при его выполнении *должна сохраняться последовательность символов* в строке и *не должно быть перестановок*.

В простейшем случае выравниваются две последовательности – парное выравнивание (англ. *pair sequence alignment*). В более сложных случаях выравнивается целый набор последовательностей – множественное выравнивание (англ. *multiple sequence alignment*, MSA) [9]. Цель парного (или «попарного») выравнивания последовательностей – идентификация областей сходства путем оптимального расположения двух последовательностей друг относительно друга. Как правило, MSA осуществляется на основе результатов парного выравнивания, часто путем слияния парных выравниваний для всех последовательностей. Последняя строка, показывающая символы, сохраненные во всех последовательностях выравнивания, называется консенсусной последовательностью. Последовательности могут быть выровнены по всей их длине (глобальное выравнивание, англ. *global alignment*) или только по отдельным участкам (локальное выравнивание, англ. *local alignment)*. Для решения отдельных задач используются и другие варианты выравнивания, например, перекрывающееся (англ. *overlap*) и точечное (англ. *dot plot*). [10–13].

Получаемая в ходе MSA консенсусная последовательность представляет собой абстрактную последовательность, которая позволяет наглядно продемонстрировать степень консервативности выравниваемых последовательностей с помощью специальных *консенсусных символов* (англ. *consensus symbols*):

- «*» (звездочка) указывает на позиции, которые имеют один полностью консервативный остаток, и означает, что остатки или нуклеотиды идентичны во всех последовательностях в выравнивании;
- «:» (двоеточие) указывает на консервацию между группами сильно схожих свойств и означает, что наблюдаются консервативные замены;
- «.» (точка) указывает на консервацию между группами слабо схожих свойств и означает, что наблюдаются полуконсервативные замены, т. е. аминокислоты, имеющие схожую форму;
- «?» (вопросительный знак) указывает на отсутствие заметного сходства в остатках в этой позиции.

Эти символы могут различаться при разных методах ее получения и визуализации.

Степень консервации зависит как от *частоты остатков* в соответствующей позиции сравниваемых последовательностей, так и от *сходства характеристик заменяемых остатков* (если они есть). Высокая степень консервации остатков в исследуемом локусе свидетельствует о его критической роли в поддержании третичной структуры или функции белка, а варианты (мутации), возникающие в неизменных, консервативных позициях, с высокой вероятностью являются причиной наследственной патологии [14].

Для каждого возможного способа сопоставления последовательностей рассчитывают специальную оценку сходства – счет выравнивания (англ. *alignment score*). Этот счет представляет сумму оценок (баллов) по каждой сравниваемой позиции в последовательностях. Алгоритм начисляет положительные баллы за каждое совпадение (англ. *match*) символов, тогда как несовпадения (англ. *mismatch*) снижают итоговый счет или имеют отрицательный вес. Любая попытка искусственно растянуть последовательность путем вставки пропуска (гэпа, англ. *gap*) карается штрафом, что гарантирует поиск биологически оправданного, а не случайного сходства. Как правило, баллы устанавливаются заранее, например, match = +2, mismatch = -1, gap = -2. Оптимальным выравниванием называют такое, которое путем перебора всех возможных вариантов сопоставлений имеет максимальный счет и соответствует биологическим закономерностям [15].

Наиболее известными алгоритмами для глобального и локального выравнивания являются: алгоритм Нидлмана – Вунша [16] для глобального выравнивания и алгоритм Смита – Уотермана [17] для решения задачи локального выравнивания. Подробное обсуждение множества методов и алгоритмов выравнивания последовательностей выходит за рамки этой работы.

Чтобы провести парное выравнивание нуклеотидных или аминокислотных последовательностей, необходимо указать сравниваемые последовательности, описать систему назначения баллов в счет выравнивания (за совпадения/несовпадения остатков, штрафы за замены или пропуски и пр.), а также специальные параметры алгоритма выравнивания.

В задачах биоинформатического выравнивания обычно выделяют два типа последовательностей: pattern (шаблон, запрос) – это искомая последовательность (или несколько последовательностей), представляющая интерес, и subject (субъект, цель) – последовательность, в которой происходит поиск (например, участок генома).

В среде R последовательности для выравнивания импортируются из различных форматов файлов (*.fasta, *.gtf и т. п.) в объекты, доступные для обработки биоинформационными пакетами [18].

Сопоставляемые последовательности могут иметь неравную длину или иметь различия, несовпадения в элементах последовательности (основаниях), вызванные мутациями, делециями или вставками в процессе эволюции. Для компенсации различий в длине выравниваемых последовательностей используется пропуск (англ. *gap*). Пропуск в выравнивании последовательностей нуклеиновых кислот или белков обозначает вставку символа пробела («-») в более короткую последовательность относительно другой.

Несовпадение в элементах последовательности (нуклеотидов или аминокислот) при выравнивании оценивается по-разному. Оно может быть приемлемым при определенном, например, биохимическом сходстве сопоставляемых нуклеотидов или аминокислот. Степень приемлемости замены (англ. *substitution*) представляется в виде матрицы весовых коэффициентов или *матрицы замен* для любой возможной пары замен нуклеотида (или аминокислоты) *i* на нуклеотид (или аминокислоту) *j* [19]. Чем выше вероятность замены, тем больше вес. Например, аминокислоты с близкими биохимическими свойствами (заряд, полярность и т. д.) чаще замещают друг друга в процессе эволюции, а другие, например, цистеин, глицин, триптофан, заменяются очень редко. Весовые коэффициенты рассчитываются на основе статистического анализа средних значений частот аминокислотных замен, выполненных на обширных наборах данных. Чтобы избежать повторов и сократить размер матриц замен в них, кроме стандартных букв, обозначающих нуклеотидные остатки, используются 11 подстановочных кодов, знаков – символов «неоднозначности» (англ. *ambiguity*), которые соответствуют одновременно нескольким возможным комбинациям четырех оснований ДНК.

Для оценки результатов выравнивания последовательностей могут использоваться общие или сводные показатели выравнивания (счет выравнивания, процент идентичности сравниваемых последовательностей).

Для вычислительных задач с последовательностями использовались пакеты программы с открытым исходным кодом на языке R, применяемом для сложных задач [20–23].

Для демонстрации возможностей выравнивания использованы данные нуклеотидной последовательности гена *GJB2*, кодирующего аминокислотную последовательность белка коннексина 26 (Cx26). Самый частый вариант, приводящий к наследственной глухоте в гене *GJB2* (белок коннексин 26) – это мутация 35delG на уровне ДНК. Но если речь идет о замене аминокислот (миссенс-вариантах), распространенными являются варианты W24C или V37I (валин в 37-й позиции заменен на изолейцин). Для проведения филогенетического анализа и оценки консервативности была сформирована выборка из девяти ортологичных последовательностей белка коннексин 26 (англ. *Connexin 26*, Entry Name: CXB2), охватывающая различные таксономические группы: приматов (*H. sapiens*, *G. gorilla*, *P. pygmaeus*, *M. mulatta*, *H. lar*), грызунов (*M. musculus*, *R. norvegicus*) и парнокопытных (*B. taurus*, *O. aries*). Использование ортологов из разных отрядов млекопитающих позволяет верифицировать функциональную значимость локуса V37 в широком эволюционном масштабе. Эти последовательности были скачаны и сохранены в виде fasta-файла (под именем «CXB2_uniprot.fasta») с сайта UniProt[1] [24]

Для демонстрации парного выравнивания на предмет поиска вариантов («опечаток») в аминокислотной последовательности у возможного пациента референсная последовательность этого белка программно была изменена в 37-й позиции с «V» (валин) на «I» (изолейцин).

Для выравнивания использовались пакеты R репозитория Bioconductor (Biostrings, pwalign, msa) [25]. Пакет Biostrings дает возможность импортировать (или создавать) последовательности в универсальные R-объекты, к которым применимы различные функции других R-пакетов. Пакет pwalign содержит основные функции, используемые при выравнивании – pairwiseAlignment() [26]. Пакет msa предоставляет унифицированный интерфейс R/ Bioconductor для широко известных алгоритмов MSA последовательностей ClustalW, ClustalOmega и Muscle. Алгоритмы MSA последовательностей дополняются функцией msaPrettyPrint() для «красивой» (англ. *pretty*) печати MSA последовательностей с использованием пакета LaTeX TeXshade [27].

РЕЗУЛЬТАТЫ

На примере белковых последовательностей коннексина 26 разных видов мы демонстрируем ключевые методы биоинформатического анализа белков. Работа с этими последовательностями позволяет отработать основные алгоритмы обработки биологических данных: поиск областей локального сходства, поиск общности происхождения, получение консенсусной последовательности, установление эволюционного родства. Здесь последовательно применяются различные методы выравнивания с использованием пакетов Biostrings, msa и pwalign в R [28, 29].

В приведенном ниже R-коде реализованы следующие этапы: 1) загрузка данных из fasta-файла, вывод на экран последовательностей и их имен (*рис. 1*); 2) создание демонстрационной измененной последовательности пациента (на базе эталонной последовательности, но с мутацией V37I); 3) запуск MSA для набора последовательностей белка коннексина 26 у разных животных; 4) получение и вывод на экран результатов выравнивания, степени консервации (*рис. 2*); 5) вывод «красивого» графика MSA (*рис. 3*); 6) сравнение последовательности коннексина 26 у пациента с эталоном человека (парное выравнивание), обнаружение варианта мутации V37I; 7) вывод статистики парного выравнивания (PID, score) и обнаруженный вариант в 37-й позиции (*рис. 4*).

Для получения качественно оформленного и визуально привлекательного вывода выравнивания аминокислотных

[1] https://www.uniprot.org.

Рисунок 1. Загрузка данных из fasta-файла, вывод на экран последовательностей белка коннексина 26 и их имен в программной среде R

Figure 1. Importing data from a FASTA file: displaying Connexin 26 protein sequences and their corresponding identifiers within the R environment

```
> library(Biostrings)
> library(msa)
> library(bio3d)
>
> # 1. Загрузка данных из fasta-файла.
> fasta_data <- readAAStringSet("CXB2_uniprot.fasta")
> # Вывод на экран всех последовательностей:
> fasta_data
AAStringSet object of length 9:
    width seq                                               names
[1]   226 MDWGTLQSILGGVNKHSTSIGKI...LNITELCYLFIRYCSGKSKRPV sp|P21994|CXB2_RA...
[2]   226 MDWGTLQTILGGVNKHSTSIGKI...LNVTELCYLLIRYCSGKSKKPV sp|P29033|CXB2_HU...
[3]   226 MDWGTLQSILGGVNKHSTSIGKI...LNITELCYLFVRYCSGKSKRPV sp|Q00977|CXB2_MO...
[4]   226 MDWGGLHTILGGVNKHSTSIGKI...LNVTELCYLLIRFCSGKSKKPV sp|A2VE67|CXB2_BO...
[5]   226 MDWSALQTILGGVNKHSTSIGKI...LNVTELCYLLIRFCSGKSKKPV sp|P46691|CXB2_SH...
[6]   226 MDWGTLQTILGGVNKHSTSIGKI...LNVTELCYLLIRYCSGKSKKPV sp|Q7JGL3|CXB2_HY...
[7]   226 MDWGTLQTILGGVNKHSTSIGKI...LNVTELCYLLIRYCSGKSKKPV sp|Q8MHW5|CXB2_GO...
[8]   226 MDWGALQTILGGVNKYSTSIGKI...LNVTELCYLLIRYCSGKSKKPV sp|Q8MIT8|CXB2_MA...
[9]   226 MDWGALQTILGGVNKHSTSIGKI...LNVTELCYLLIRYCSGRSKKPV sp|Q8MIT9|CXB2_PO...
> # Список имен видов(кратко):
> names(fasta_data) <- sub(".*OS=(.*) OX=.*", "\\1", names(fasta_data))
> print(names(fasta_data))
[1] "Rattus norvegicus"       "Homo sapiens"
[3] "Mus musculus"            "Bos taurus"
[5] "Ovis aries"              "Hylobates lar"
[7] "Gorilla gorilla gorilla" "Macaca mulatta"
[9] "Pongo pygmaeus"
```

Рисунок 2. Процесс создания измененной последовательности пациента с мутацией V37I и последующее множественное выравнивание (MSA) с последовательностями коннексина 26 у разных организмов, отражающее уровни гомологии и консервативности

Figure 2. Generating the V37I mutant sequence and performing multiple sequence alignment (MSA) with Connexin 26 orthologs to illustrate homology levels and evolutionary conservation

```
> # 2. Создание демонстрационной измененной последовательности пациента (на базе
эталонной последовательности, но с мутацией V37I)
> human_ref <- fasta_data["Homo sapiens"]
> patient_seq <- human_ref
> subseq(patient_seq, start=37, end=37) <- "I"
> names(patient_seq) <- "Patient (V37I)"
> patient_seq
AAStringSet object of length 1:
    width seq                                               names
[1]   226 MDWGTLQTILGGVNKHSTSIGKI...LNVTELCYLLIRYCSGKSKKPV Patient (V37I)
>
> # 3. Запуск множественного выравнивания для набора последовательностей белка
коннексина-26 разных животных
> full_msa <- msa(fasta_data)
use default substitution matrix
>
> # 4. Получение и вывод на экран результатов выравнивания, степени консервации.
> msa_bio3d <- msaConvert(full_msa, type="bio3d::fasta")
> cons <- conserv(msa_bio3d$ali, method="identity")
> # Вывод на экран индекса именно в 37-й позиции
> print(paste("Conservation Score at pos 37:", cons[37]))
[1] "Conservation Score at pos 37: 1"
```

последовательностей в формате LaTeX (*рис. 3*) использовался пакет msa с функцией msaPrettyPrint(). Данная функция позволила представить результаты выравнивания в виде эстетичной таблицы, подходящей для публикации и отчетов. Чтобы обеспечить поддержку LaTeX в среде R, дополнительно установлен и настроен TinyTeX – легковесный и полный комплект инструментов для работы с LaTeX.

Результаты MSA подтвердили абсолютную инвариантность валина в 37-й позиции белка *GJB2* во всех исследованных филогенетических группах. При интеграции данных пациента с вариантом V37I в общий массив индекс консервативности в данной позиции составил менее 1.0, что при полной гомологии остальных сайтов фрагмента (30–45-й позиций аминокислотной последовательности) свидетельствует о патогенном потенциале замены. В дальнейшем, если использовать критерии патогенности Американского колледжа медицинской генетики и Ассоциации молекулярной патологии (ACMG/AMP) [30]) для клинической интерпретации этого результата, он объективизирует присвоение критерия PP3 (см. раздел «Обсуждение»). Если парное выравнивание лишь фиксирует факт замены (V37I), то MSA позволяет оценить ее биологический вес.

Этот пример наглядно демонстрирует пошаговое применение биоинформатических методов для сравнительного исследования белковых последовательностей, выявления консервативных доменов и анализа эволюционных взаимоотношений между видами. Полученные результаты имеют важное значение для понимания структурно-функциональных особенностей коннексина 26 и молекулярных механизмов, лежащих в основе нарушений слуха.

ОБСУЖДЕНИЕ

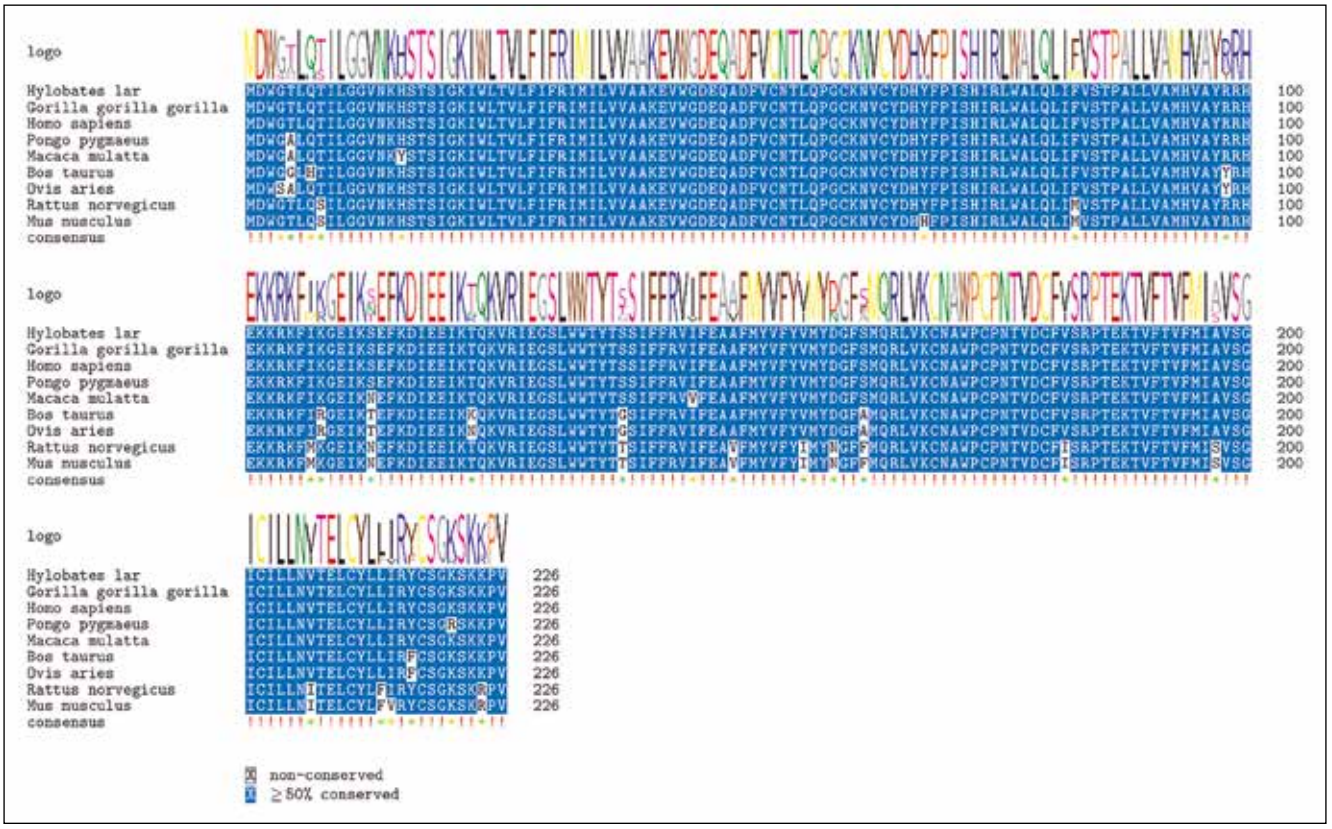
В данной работе рассмотрены фундаментальные аспекты выравнивания биологических последовательностей и их прикладное значение для современной медицинской биоинформатики. Методологические подходы, описанные в статье, являются фундаментом для анализа генетических данных, что критически важно для развития персонализированной медицины и точной диагностики наследственных заболеваний.

Связующим звеном между программно-техническим анализом и постановкой диагноза является интерпретация данных выравнивания в рамках международных стандартов. Ключевым инструментом здесь выступают рекомендации ACMG/AMP – общепризнанный мировой стандарт классификации вариантов.

Система буквенно-цифровых кодов ACMG позволяет врачу объективно разграничить патогенные мутации и нейтральные генетические особенности (доброкачественные варианты). В этом процессе выравнивание играет роль «эволюционного фильтра»: высокая степень

● ***Рисунок 3.*** Результат визуализации множественного выравнивания (MSA) аминокислотных последовательностей, выполненного с использованием функции `msaPrettyPrint()` пакета msa

● ***Figure 3.*** Visualizing the multiple amino acid sequence alignment generated using the msaPrettyPrint() function from the msa package



консервации остатка в определенной позиции свидетельствует о его функциональной незаменимости. Обнаружение замены в такой инвариантной точке позволяет исследователю применить прогностический критерий PP3, подтверждающий патогенный потенциал находки.

Использование различных типов выравнивания обеспечивает многоуровневый анализ данных: парное выравнивание эффективно для локального поиска и идентификации конкретных «опечаток», тогда как MSA предоставляет возможности сравнительной геномики, позволяя выявлять консервативные домены и филогенетические паттерны.

ЗАКЛЮЧЕНИЕ

Практический пример анализа коннексина 26 наглядно демонстрирует, как биоинформатические методы выявляют структурно-функциональные аномалии белков, которые лежат в основе патогенеза наследственных заболеваний.

● ***Рисунок 4.*** Результат парного выравнивания аминокислотной последовательности белка коннексина 26, иллюстрирующий наличие мутации V37I и уровень сходства последовательностей (PID, score)

● ***Figure 4.*** Pairwise amino acid sequence alignment of Connexin 26, highlighting the V37I mutation and sequence metrics, including Percent Identity (PID) and alignment score

```
> # 6. Сравнение последовательности пациента с эталонной (парное выравнивание),
обнаружение мутации V37I
> pair_align <- pwalign::pairwiseAlignment(patient_seq, fasta_data["Homo sapiens"])
> # 7. Вывод статистики парного выравнивания (PID, score) и обнаруженного варианта в
37 позиции.
> #  Общие статистики, включая несоответствия, число соответствий и счет выравнивания:
> summary(pair_align)
Global Single Subject Pairwise Alignments
Number of Alignments:  1

Scores:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  962.1   962.1   962.1   962.1   962.1   962.1

Number of matches:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    225     225     225     225     225     225

Top 1 Mismatch Counts:
  SubjectPosition Subject Pattern Count Probability
1              37       V       I     1           1
> print(pair_align)
Global PairwiseAlignmentsSingleSubject (1 of 1)
pattern: MDWGTLQTILGGVNKHSTSIGKIWLTVLFIFRIM...FMIAVSGICILLNVTELCYLLIRYCSGKSKKPV
subject: MDWGTLQTILGGVNKHSTSIGKIWLTVLFIFRIM...FMIAVSGICILLNVTELCYLLIRYCSGKSKKPV
score: 962.085
> # Вывод процента идентичности:
> pid_value <- pwalign::pid(pair_align)
> print(pid_value)
[1] 99.55752
```

Особое внимание стоит уделить инструментарию: использование языка R и экосистемы Bioconductor упрощает изучение и использование биоинформатических методов. Представленные в статье алгоритмы и примеры кода доказывают, что современные методы анализа доступны врачам и исследователям без углубленной вычислительной подготовки. Это способствует демократизации геномных технологий и их оперативному внедрению в широкую клиническую практику.

Поступила / Received 18.02.2026
Поступила после рецензирования / Revised 12.03.2026
Принята в печать / Accepted 12.03.2026

Список литературы / References

1. Корнеенков АА, Янов ЮК, Вяземская ЕЭ, Медведева АЮ. От данных секвенирования к пониманию болезни: как врачу обработать NGS-данные пациента на своем компьютере. *Медицинский совет*. 2025;19(18):108–121. (In Russ.) https://doi.org/10.21518/ms2025-351.
Korneenkov AA, Yanov YuK, Vyazemskaya EE, Medvedeva AYu. From sequencing data to disease understanding: How can a doctor process patient's NGS data on their own computer. *Meditsinskiy Sovet*. 2025;19(18):108–121. (In Russ.) https://doi.org/10.21518/ms2025-351.
2. Bawono P, Dijkstra M, Pirovano W, Feenstra A, Abeln S, Heringa J. Multiple Sequence Alignment. In: Keith JM (ed.). *Bioinformatics: Volume I: Data, Sequence Analysis, and Evolution*. Springer; 2017, pp. 167–189. https://doi.org/10.1007/978-1-4939-6622-6_8.
3. Корнеенков АА, Янов ЮК, Вяземская ЕЭ, Медведева АЮ. Вопросы интеграции медицинских биоинформатических технологий в оториноларингологию: проблемы и программные решения. *Российская оториноларингология*. 2024;23(6):8–19. https://doi.org/10.18692/1810-4800-2024-6-8-19.
Korneenkov AA, Yanov YuK, Vyazemskaya EE, Medvedeva AYu. Issues of integrating medical bioinformatics technologies into otorhinolaryngology: challenges and software solutions. *Rossiiskaya Otorinolaringologiya*. 2024;23(6):8–19. (In Russ.) https://doi.org/10.18692/1810-4800-2024-6-8-19.
4. Sabonsolin J, Lao D. A comprehensive systematic literature review of multiple sequence alignment algorithms. *Discov Computing*. 2026;29:33. https://doi.org/10.1007/s10791-026-09911-3.
5. Каляпин ДД, Сугарова СБ, Кузовков ВЕ, Лиленко АС, Преображенская ЮС. Этиологический спектр врожденной глухоты и его значение в кохлеарной имплантации. *Российская оториноларингология*. 2019;18(1):41–45. https://doi.org/10.18692/1810-4800-2019-1-41-45.
Kalyapin DD, Sugarova SB, Kuzovkov VE, Lilenko AS, Preobrazhenskaya YuS. Congenital deafness etiologic spectrum and its importance in cochlear implantation. *Rossiiskaya Otorinolaringologiya*. 2019;18(1):41–45. (In Russ.) https://doi.org/10.18692/1810-4800-2019-1-41-45.
6. Кузовков ВЕ, Сугарова СБ, Лиленко АС, Преображенская ЮС, Каляпин ДД, Скирпичников ИН. Особенности хирургического этапа кохлеарной имплантации у пациентов с ЦМВ и GJB2-ассоциированной глухотой. *Российская оториноларингология*. 2020;19(4):55–59. https://doi.org/10.18692/1810-4800-2020-4-55-59.
Kuzovkov VE, Sugarova SB, Lilenko AS, Preobrazhenskaya YuS, Kalyapin DD, Skirpichnikov IN. Features of the surgical stage of cochlear implantation in patients with CMV and GJB2-associated deafness. *Rossiiskaya Otorinolaringologiya*. 2020;19(4):55–59. (In Russ.) https://doi.org/10.18692/1810-4800-2020-4-55-59.
7. Маркова ТГ, Маркова МВ. Знание о генетических причинах тугоухости – ключ к профилактике. *Сенсорные системы*. 2025;39(4):5–20. https://doi.org/10.7868/S3034593625040019.
Markova TG, Markova MV. Knowledge of the Genetic Causes of Hearing Loss is the Key to its Prevention. *Sensory Systems*. 2025;39(4):5–20. (In Russ.) https://doi.org/10.7868/S3034593625040019.
8. Халфина ВВ, Степанова АА, Маркова ТГ, Поляков АВ, Тавартkиладзе ГА, Насыров ВА. Мутации в гене GJB2 у детей с двусторонней тугоухостью в Кыргызстане. *Российская оториноларингология*. 2020;19(6):64–71. https://doi.org/10.18692/1810-4800-2020-6-64-71.
Khalfina VV., Stepanova AA, Markova TG, Polyakov AV, Tavartkiladze GA, Nasyrov VA. Mutations in the gjb2 gene in children with bilateral hearing loss in Kyrgyzstan. *Rossiiskaya Otorinolaringologiya*. 2020;19(6):64–71. (In Russ.) https://doi.org/10.18692/1810-4800-2020-6-64-71.
9. Wang Z, Tan J, Long Y, Liu Y, Lei W, Cai J et al. SaAlign: Multiple DNA/RNA sequence alignment and phylogenetic tree construction tool for ultra-large datasets and ultra-long sequences based on suffix array. *Comput Struct Biotechnol J*. 2022;20:1641–1652. https://doi.org/10.1016/j.csbj.2022.03.018.
10. Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ*. 2018;6:e4958. https://doi.org/10.7717/peerj.4958.
11. Firth AE, Brown CM. Detecting overlapping coding sequences with pairwise alignments. *Bioinformatics*. 2005;21(3):282–292. https://doi.org/10.1093/bioinformatics/bti007.
12. Chu J, Mohamadi H, Warren RL, Yang C, Birol I. Innovations and challenges in detecting long read overlaps: an evaluation of the state-of-the-art. *Bioinformatics*. 2017;33(8):1261–1270. https://doi.org/10.1093/bioinformatics/btw811.
13. Kirzhner V, Frenkel S, Korol A. Minimal-dot plot: "Old tale in new skin" about sequence comparison. *Inf Sci*. 2011;181(8):1454–1462. https://doi.org/10.1016/j.ins.2010.12.009.
14. Chesneau B, Aubert-Mucca M, Fremont F, Pechmeja J, Soler V, Isidor B et al. First evidence of SOX2 mutations in Peters'anomaly: Lessons from molecular screening of 95 patients. *Clin Genet*. 2022;101(5-6):494–506. https://doi.org/10.1111/cge.14123.
15. Bahk K, Sung J. SigAlign: an alignment algorithm guided by explicit similarity criteria. *Nucleic Acids Res*. 2024;52(15):8717–8733. https://doi.org/10.1093/nar/gkae607.
16. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*. 1970;48(3):443–453. https://doi.org/10.1016/0022-2836(70)90057-4.
17. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol*. 1981;147(1):195–197. https://doi.org/10.1016/0022-2836(81)90087-5.
18. Корнеенков АА, Янов ЮК, Дворянчиков ВВ, Вяземская ЕЭ, Медведева АЮ. Медицинская биоинформатика: базовые операции с нуклеотидными последовательностями в программной среде R. *Российская оториноларингология*. 2025;24(4):13–27. https://doi.org/10.18692/1810-4800-2025-4-13-27.
Korneenkov AA, Yanov YuK, Dvoryanchikov VV, Vyazemskaya EE, Medvedeva AYu. Medical bioinformatics: basic operations with nucleotide sequences in R software environment. *Rossiiskaya Otorinolaringologiya*. 2025;24(4):13–27. (In Russ.) https://doi.org/10.18692/1810-4800-2025-4-13-27.
19. Trivedi R, Nagarajaram HA. Substitution scoring matrices for proteins – An overview. *Protein Sci*. 2020;29(11):2150–2163. https://doi.org/10.1002/pro.3954.
20. Корнеенков АА, Рязанцев СВ, Левин СВ, Храмов АВ, Вяземская ЕЭ, Скирпичников ИН и др. Пространственно-статистический анализ данных о нарушениях слуха у жителей Челябинской области. *Российская оториноларингология*. 2021;20(3):39–50. https://doi.org/10.18692/1810-4800-2021-3-39-50.
Korneenkov AA, Ryazantsev SV, Levin SV, Khramov AV, Vyazemskaya EE, Skirpichnikov IN et al. Spatial and statistical analysis of hearing impairment data of Chelyabinsk region residents. *Rossiiskaya Otorinolaringologiya*. 2021;20(3):39–50. (In Russ.) https://doi.org/10.18692/1810-4800-2021-3-39-50.
21. Корнеенков АА, Янов ЮК, Рязанцев СВ, Вяземская ЕЭ, Астащенко СВ, Рязанцева ЕС. Метаанализ клинических исследований в оториноларингологии. *Вестник оториноларингологии*. 2020;85(2):26–30. https://doi.org/10.17116/otorino20208502126.
Korneenkov AA, Yanov YuK, Ryazantsev SV, Vyazemskaya EE, Astashchenko SV, Ryazantseva ES. A meta-analysis of clinical studies in otorhinolaryngology. *Vestnik Otorinolaringologii*. 2020;85(2):26–30. (In Russ.) https://doi.org/10.17116/otorino20208502126.
22. Корнеенков АА, Левина ЕА, Вяземская ЕЭ, Левин СВ, Скирпичников ИН. Пространственный кластерный анализ в моделировании доступности медицинской помощи пожилым пациентам с нарушениями слуха. *Российская оториноларингология*. 2021;20(6):8–19. https://doi.org/10.18692/1810-4800-2021-6-8-19.
Korneenkov AA, Levina EA, Vyazemskaya EE, Levin SV, Skirpichnikov IN. Spatial cluster modeling of access of elderly patients with hearing loss to medical services. *Rossiiskaya Otorinolaringologiya*. 2021;20(6):8–19. (In Russ.) https://doi.org/10.18692/1810-4800-2021-6-8-19.
23. Корнеенков АА, Рязанцев СВ, Вяземская ЕЭ, Будковая МА. Меры информативности диагностических медицинских технологий в оториноларингологии: вычисление и интерпретация. *Российская оториноларингология*. 2020;19(1):46–55. https://doi.org/10.18692/1810-4800-2020-1-46-55.
Korneenkov AA, Ryazantsev SV, Vyazemskaya EE, Budkovaya MA. The measures of informativeness of diagnostic medical technologies in otorhinolaryngology: calculation and interpretation. *Rossiiskaya Otorinolaringologiya*. 2020;19(1):46–55. (In Russ.) https://doi.org/10.18692/1810-4800-2020-1-46-55.
24. Буг ДС, Наркевич АН, Петухова НВ. Обзор программ для оценки патогенности генетических вариантов. *Ученые записки Первого Санкт-Петербургского государственного медицинского университета имени академика И.П. Павлова*. 2025;32(1):11–20. https://doi.org/10.24884/1607-4181-2025-32-1-11-20.

Bug DS, Narkevich AN, Petukhova NV. Review of programs for assessing the pathogenicity of genetic variants. *The Scientific Notes of the Pavlov University.* 2025;32(1):11–20. (In Russ.) https://doi.org/10.24884/1607-4181-2025-32-1-11-20.
25. Toparslan E, Karabag K, Bilge U. A workflow with R: Phylogenetic analyses and visualizations using mitochondrial cytochrome b gene sequences. *PLoS ONE.* 2020;15(12):e0243927. https://doi.org/10.1371/journal.pone.0243927.
26. Aboyoun P, Gentleman R. *pwalign: Perform pairwise sequence alignments.* R package version 1.6.0. 2025. https://doi.org/10.18129/B9.bioc.pwalign.
27. Bodenhofer U, Bonatesta E, Horejs-Kainrath C, Hochreiter S. msa: an R package for multiple sequence alignment. *Bioinformatics.* 2015;31(24):3997–3999. https://doi.org/10.1093/bioinformatics/btv494.
28. Корнеенков АА, Вяземская ЕЭ, Медведева АЮ, Янов ЮК. *Программа подготовки тренировочных наборов нуклеотидных последовательностей для изучения методов парного выравнивания в образовательной практике биоинформатики.* Свидетельство о государственной регистрации программы для ЭВМ RU 2025668584, 07.07.2025. Режим доступа: https://elibrary.ru/kbtffy.
29. Корнеенков АА, Косачев КА, Медведева АЮ, Вяземская ЕЭ. *Интерактивное учебное пособие tbioinf «Введение в биоинформатику. Выравнивание генетических последовательностей и построение филогенетических деревьев в R» (теория, визуализация, практика, тестирование).* Свидетельство о государственной регистрации программы для ЭВМ RU 2025683571, 19.08.2025. Режим доступа: https://elibrary.ru/xrjmbo.
30. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med.* 2015;17(5):405–424. https://doi.org/10.1038/gim.2015.30.

Вклад авторов:

Концепция статьи – **А.А. Корнеенков**
Концепция и дизайн исследования – **А.А. Корнеенков**
Написание текста – **А.А. Корнеенков**
Сбор и обработка материала – **Е.Э. Вяземская, А.Ю. Медведева**
Обзор литературы – **А.А. Корнеенков**
Анализ материала – **А.А. Корнеенков**
Статистическая обработка – **А.А. Корнеенков, Е.Э. Вяземская, А.Ю. Медведева**
Редактирование – **Ю.К. Янов, Е.Э. Вяземская**
Утверждение окончательного варианта статьи – **Ю.К. Янов, А.А. Корнеенков**

Contribution of authors:

Concept of the article – **Aleksei A. Korneenkov**
Study concept and design – **Aleksei A. Korneenkov**
Text development – **Aleksei A. Korneenkov**
Collection and processing of material – **Elena E. Vyazemskaya, Anna Yu. Medvedeva**
Literature review – **Aleksei A. Korneenkov**
Material analysis – **Aleksei A. Korneenkov**
Statistical processing – **Aleksei A. Korneenkov, Elena E. Vyazemskaya, Anna Yu. Medvedeva**
Editing – **Yuri K. Yanov, Elena E. Vyazemskaya**
Approval of the final version of the article – **Yuri K. Yanov, Aleksei A. Korneenkov**

Информация об авторах:

Корнеенков Алексей Александрович, д.м.н., профессор, заведующий научно-исследовательской лабораторией клинической информатики и биостатистики, Санкт-Петербургский научно-исследовательский институт уха, горла, носа и речи; 190013, Россия, Санкт-Петербург, ул. Бронницкая, д. 9; korneyenkov@gmail.com
Янов Юрий Константинович, академик РАН, д.м.н., профессор, профессор кафедры оториноларингологии, Военно-медицинская академия имени С.М. Кирова; 194044, Россия, Санкт-Петербург, ул. Академика Лебедева, д. 6; профессор кафедры оториноларингологии, Северо-Западный государственный медицинский университет имени И.И. Мечникова; 191015, Россия, Санкт-Петербург, ул. Кирочная, д. 41; 9153764@mail.ru
Вяземская Елена Эмильевна, младший научный сотрудник, инженер научно-исследовательской лаборатории клинической информатики и биостатистики, Санкт-Петербургский научно-исследовательский институт уха, горла, носа и речи; 190013, Россия, Санкт-Петербург, ул. Бронницкая, д. 9; vyazemskaya.elena@gmail.com
Медведева Анна Юрьевна, младший научный сотрудник, инженер научно-исследовательской лаборатории клинической информатики и биостатистики, Санкт-Петербургский научно-исследовательский институт уха, горла, носа и речи; 190013, Россия, Санкт-Петербург, ул. Бронницкая, д. 9; a.medvedeva@niilor.ru

Information about the authors:

Aleksei A. Korneenkov, Dr. Sci. (Med.), Professor, Head of the Research Laboratory of Clinical Informatics and Biostatistics, Saint Petersburg Research Institute of Ear, Throat, Nose and Speech; 9, Bronnitskaya St., St Petersburg, 190013, Russia; korneyenkov@gmail.com
Yuri K. Yanov, Acad. RAS, Dr. Sci. (Med.), Professor, Professor of the Department of Otolaryngology, Kirov Military Medical Academy; 6, Akademik Lebedev St., St Petersburg, 194044, Russia; Professor of the Department of Otolaryngology, North-Western State Medical University named after I.I. Mechnikov; 41, Kirochnaya St., St Petersburg, 191015, Russia; 9153764@mail.ru
Elena E. Vyazemskaya, Junior Researcher, Engineer of the Research Laboratory of Clinical Informatics and Biostatistics, Saint Petersburg Research Institute of Ear, Throat, Nose and Speech; 9, Bronnitskaya St., St Petersburg, 190013, Russia; vyazemskaya.elena@gmail.com
Anna Yu. Medvedeva, Junior Researcher, Engineer of the Research Laboratory of Clinical Informatics and Biostatistics, Saint Petersburg Research Institute of Ear, Throat, Nose and Speech; 9, Bronnitskaya St., St Petersburg, 190013, Russia; a.medvedeva@niilor.ru