ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

«ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
ҰЛТТЫҚ ҒЫЛЫМ
АКАДЕМИЯСЫ» РҚБ

ҚАЙЫРЫМДЫЛЫҚ ҚОРЫ
HALYK
CHARITY FOUNDATION

«ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
ҰЛТТЫҚ ҒЫЛЫМ АКАДЕМИЯСЫ» РҚБ
«ХАЛЫҚ» ЖҚ

ХАБАРЛАРЫ

ИЗВЕСТИЯ

РОО «НАЦИОНАЛЬНОЙ
АКАДЕМИИ НАУК РЕСПУБЛИКИ
КАЗАХСТАН»
ЧФ «Халық»

N E W S

OF THE ACADEMY OF SCIENCES
OF THE REPUBLIC OF
KAZAKHSTAN
«Halyk» Private Foundation

SERIES
PHYSICS AND INFORMATION TECHNOLOGY

3 (347)

JULY – SEPTEMBER 2023

PUBLISHED SINCE JANUARY 1963
PUBLISHED 4 TIMES A YEAR

ALMATY, NAS RK

ЧФ «ХАЛЫҚ»

В 2016 году для развития и улучшения качества жизни казахстанцев был создан частный Благотворительный фонд «Халык». За годы своей деятельности на реализацию благотворительных проектов в областях образования и науки, социальной защиты, культуры, здравоохранения и спорта, Фонд выделил более 45 миллиардов тенге.

Особое внимание Благотворительный фонд «Халык» уделяет образовательным программам, считая это направление одним из ключевых в своей деятельности. Оказывая поддержку отечественному образованию, Фонд вносит свой посильный вклад в развитие качественного образования в Казахстане. Тем самым способствуя росту числа людей, способных менять жизнь в стране к лучшему – профессионалов в различных сферах, потенциальных лидеров и «великих умов». Одной из значимых инициатив фонда «Халык» в образовательной сфере стал проект Ozgeris powered by Halyk Fund – первый в стране бизнес-инкубатор для учащихся 9-11 классов, который помогает развивать необходимые в современном мире предпринимательские навыки. Так, на содействие малому бизнесу школьников было выделено более 200 грантов. Для поддержки талантливых и мотивированных детей Фонд неоднократно выделял гранты на обучение в Международной школе «Мирас» и в Astana IT University, а также помог казахстанским школьникам принять участие в престижном конкурсе «USTEM Robotics» в США. Авторские работы в рамках проекта «Тәлімгер», которому Фонд оказал поддержку, легли в основу учебной программы, учебников и учебно-методических книг по предмету «Основы предпринимательства и бизнеса», преподаваемого в 10-11 классах казахстанских школ и колледжей.

Помимо помощи школьникам, учащимся колледжей и студентам Фонд считает важным внести свой вклад в повышение квалификации педагогов, совершенствование их знаний и навыков, поскольку именно они являются проводниками знаний будущих поколений казахстанцев. При поддержке Фонда «Халык» в южной столице был организован ежегодный городской конкурс педагогов «Almaty Digital Ustaz.

Важной инициативой стал реализуемый проект по обучению основам финансовой грамотности преподавателей из восьми областей Казахстана, что должно оказать существенное влияние на воспитание финансовой грамотности и предпринимательского мышления у нового поколения граждан страны.

Необходимую помощь Фонд «Халык» оказывает и тем, кто особенно остро в ней нуждается. В рамках социальной защиты населения активно проводится работа по поддержке детей, оставшихся без родителей, детей и взрослых из социально уязвимых слоев населения, людей с ограниченными возможностями, а также обеспечению нуждающихся социальным жильем, строительству социально важных объектов, таких как детские сады, детские площадки и физкультурно-оздоровительные комплексы.

В копилку добрых дел Фонда «Халык» можно добавить оказание помощи детскому спорту, куда относится поддержка в развитии детского футбола и карате в нашей стране. Жизненно важную помощь Благотворительный фонд «Халык» оказал нашим соотечественникам во время недавней пандемии COVID-19. Тогда, в разгар тяжелой борьбы с коронавирусной инфекцией Фонд выделил свыше 11 миллиардов тенге на приобретение необходимого медицинского оборудования и дорогостоящих медицинских препаратов, автомобилей скорой медицинской помощи и средств защиты, адресную материальную помощь социально уязвимым слоям населения и денежные выплаты медицинским работникам.

В 2023 году наряду с другими проектами, нацеленными на повышение благосостояния казахстанских граждан Фонд решил уделить особое внимание науке, поскольку она является частью общественной культуры, а уровень ее развития определяет уровень развития государства.

Поддержка Фондом выпуска журналов Национальной Академии наук Республики Казахстан, которые входят в международные фонды Scopus и Wos и в которых публикуются статьи отечественных ученых, докторантов и магистрантов, а также научных сотрудников высших учебных заведений и научно-исследовательских институтов нашей страны является не менее значимым вкладом Фонда в развитие казахстанского общества.

С уважением,
Благотворительный Фонд «Халык»!

«ҚР ҰҒА Хабарлары. Физика және информатика сериясы».
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Меншіктеуші: «Қазақстан Республикасының Ұлттық ғылым академиясы» РҚБ (Алматы қ.). Қазақстан Республикасының Ақпарат және қоғамдық даму министрлігінің Ақпарат комитетінде 14.02.2018 ж. берілген **No 16906-Ж** мерзімдік басылым тіркеуіне қойылу туралы куәлік.
Тақырыптық бағыты: *физика және ақпараттық коммуникациялық технологиялар сериясы.*
Қазіргі уақытта: *«ақпараттық технологиялар»бағыты бойынша ҚР БҒМ БҒСБК ұсынған журналдар тізіміне енді.*
Мерзімділігі: *жылына 4 рет.*
Тиражы: *300 дана.*
Редакцияның мекен-жайы: *050010, Алматы қ., Шевченко көш., 28, 219 бөл., тел.: 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

«Известия НАН РК. Серия физика и информатики».
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Собственник: *Республиканское общественное объединение «Национальная академия наук Республики Казахстан» (г. Алматы).*
Свидетельство о постановке на учет периодического печатного издания в Комитете информации Министерства информации и общественного развития Республики Казахстан **No 16906-Ж** выданное 14.02.2018 г.
Тематическая направленность: *серия физика и информационные коммуникационные технологии.*
В настоящее время: *вошел в список журналов, рекомендованных ККСОН МОН РК по направлению «информационные коммуникационные технологии».*
Периодичность: *4 раз в год.*
Тираж: *300 экземпляров.*
Адрес редакции: *050010, г. Алматы, ул. Шевченко, 28, оф. 219, тел.: 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

News of the National Academy of Sciences of the Republic of Kazakhstan.
Series of physics and informatics.
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Owner: RPA «National Academy of Sciences of the Republic of Kazakhstan» (Almaty). The certificate of registration of a periodical printed publication in the Committee of information of the Ministry of Information and Social Development of the Republic of Kazakhstan **No. 16906-Ж,** issued 14.02.2018
Thematic scope: *series physics and information technology.*
Currently: *included in the list of journals recommended by the CCSES MES RK in the direction of «information and communication technologies».*
Periodicity: *4 times a year.*
Circulation: *300 copies.*
Editorial address: *28, Shevchenko str., of. 219, Almaty, 050010, tel. 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 3. Namber 347 (2023). 62–75
https://doi.org/10.32014/2023.2518-1726.204

UDC 004.057A

© Y.S. Golenko, A.A. Ismailova, 2023
NCJSC «S. Seifullin Kazakh AgroTechnical Research University»,
Astana, Kazakhstan.
E-mail: golenko.katerina@gmail.com

PROTEIN FUNCTION PREDICTION USING THE COMBINATION OF BILSTM AND SELF-ATTENTION ALGORITHM

Golenko Yekaterina Sergeyevna — Master of Science, PhD candidate, NCJSC «S. Seifullin Kazakh AgroTechnical Research University», Astana, Kazakhstan
E-mail: golenko.katerina@gmail.com; ORCID ID: https://orcid.org/0000-0002-4643-4571;
Ismailova Aisulu Abzhapparovna — PhD, associate professor, NCJSC «S. Seifullin Kazakh AgroTechnical Research University», Astana, Kazakhstan
E-mail: a.ismailova@mail.ru; ORCID ID: https://orcid.org/0000-0002-8958-1846.

Abstract. With the development of genome sequencing technology, the use of computational technologies to predict the function of proteins has become one of the important tasks of bioinformatics. Early research in this area was based on sequence similarity and assumed that proteins with similar amino acid sequences had similar functions. However, previously proposed methods for predicting functions often failed to reveal hidden patterns between proteins and gene ontology terms, which reduced the accuracy of functional annotation. Deep machine learning, as many studies show, copes with this task at a higher level. First, deep learning methods can be trained on large amounts of protein sequence data without considering additional information about protein properties. Secondly, deep learning approaches solve such side problems as data noisiness, their redundancy and high dimensionality. The combination of a self-attention mechanism and a bidirectional network with long-term short-term memory can be used to solve the problem of protein functional annotation. Bidirectional LSTM is used to obtain both global and local information about the properties of protein sequences, as well as to store the information obtained. The self-attention algorithm is applied to make optimal use of the relationship of the sequence and information about the functions of different positions of the sequence, which will increase the reliability of the prediction. The python language was chosen as a tool for implementing the algorithms, the model was trained for 50 epochs and tested on an experimental dataset of the Indica protein obtained from open sources.The results of the experiment show that the algorithm for combining

the self-attention mechanism and the bidirectional network with long-term short-term memory outperforms other traditional neural network algorithms and can more accurately predict protein function, which shows the possible applicability of the algorithm in functional annotation of protein sequences.

Keywords: bidirectional LSTM, self-attention, function prediction, proteins, machine learning

«Сәкен Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті» КеАҚ, Астана, Қазақстан.

E-mail: golenko.katerina@gmail.com

ПАЙДАЛАНУ ПРОТЕИНДІК ФУНКЦИЯЛАРЫН БОЛЖАУ BILSTM ЖӘНЕ ӨЗІН-ӨЗІ ТАНУ АЛГОРИТМІНІҢ КОМБИНАЦИЯЛАРЫ

Голенко Екатерина Сергеевна — техника ғылымдарының магистрі, PhD докторанты, «Сәкен Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті» КеАҚ, Астана, Қазақстан Республикасы
E-mail: golenko.katerina@gmail.com; ORCID ID: https://orcid.org/0000-0002-4643-4571;
Исмаилова Айсулу Абжаппаровна — PhD, ассоциированный профессор, «Сәкен Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті» КеАҚ, Астана, Қазақстан Республикасы
E-mail: a.ismailova@mail.ru; ORCID ID: https://orcid.org/0000-0002-8958-1846.

Аннотация. Геномды секвенирлеу технологиясының дамуымен белоктардың қызметін болжау үшін есептеу технологияларын қолдану биоинформатиканың маңызды міндеттерінің біріне айналды. Бұл саладағы алғашқы зерттеулер дәйектілік ұқсастығына негізделген және ұқсас аминқышқылдары тізбегі бар белоктардың ұқсас функциялары бар деп болжаған. Дегенмен, функцияларды болжау үшін бұрын ұсынылған әдістер көбінесе белоктар мен гендік онтология терминдері арасындағы жасырын заңдылықтарды аша алмады, бұл функционалдық аннотацияның дәлдігін төмендетті. Терең машиналық оқыту, көптеген зерттеулер көрсеткендей, бұл тапсырманы жоғары деңгейде жеңеді. Біріншіден, белок қасиеттері туралы қосымша ақпаратты есепке алмай-ақ, ақуыз тізбегі деректерінің үлкен көлемі бойынша терең оқыту әдістерін үйретуге болады. Екіншіден, терең оқыту тәсілдері деректердің шулылығы, олардың артықтығы және жоғары өлшемділігі сияқты жанама мәселелерді шешеді. Ұзақ мерзімді қысқа мерзімді жады бар өзіне-өзі назар аудару механизмі мен қос бағытты желі комбинациясы белоктың функционалдық аннотациясының мәселесін шешу үшін пайдаланылуы мүмкін. Екі бағытты LSTM протеин тізбегінің қасиеттері туралы ғаламдық және жергілікті ақпаратты алу үшін, сондай-ақ алынған ақпаратты сақтау үшін қолданылады. Өзіндік назар аудару алгоритмі реттілік қатынасын және реттіліктің әртүрлі позицияларының функциялары туралы ақпаратты оңтайлы пайдалану үшін қолданылады, бұл болжамның

сенімділігін арттырады. Алгоритмдерді енгізу құралы ретінде питон тілі таңдалды, модель 50 дәуір бойы оқытылды және ашық көздерден алынған Indica протеинінің тәжірибелік деректер жинағында сынақтан өтті. Тәжірибе нәтижелері көрсеткендей, өзіне-өзі назар аудару механизмі мен екі бағытты желіні ұзақ мерзімді қысқа мерзімді жадымен біріктіру алгоритмі басқа дәстүрлі нейрондық желі алгоритмдерінен асып түседі және алгоритмнің мүмкін болатын қолдану мүмкіндігін көрсететін ақуыз функциясын дәлірек болжауға болады. ақуыз тізбектерінің функционалдық аннотациясында.

Түйін сөздер: екі бағытты LSTM, өзіндік зейін, мүмкіндіктерді болжау, ақуыздар, машиналық оқыту

© Е.С. Голенко, А.А. Исмаилова, 2023
НАО «Казахский агротехнический исследовательский университет имени Сакена Сейфуллина», Астана, Казахстан.
E-mail: golenko.katerina@gmail.com

ПРЕДСКАЗАНИЕ ФУНКЦИЙ БЕЛКА С ИСПОЛЬЗОВАНИЕМ КОМБИНАЦИИ BILSTM И АЛГОРИТМА САМОВНИМАНИЯ

Голенко Екатерина Сергеевна — магистр технических наук, выпускник докторантуры, НАО «Казахский агротехнический исследовательский университет имени Сакена Сейфуллина», Астана, Казахстан
E-mail: golenko.katerina@gmail.com; ORCID ID: https://orcid.org/0000-0002-4643-4571;
Исмаилова Айсулу Абжаппаровна — PhD, ассоциированный профессор, НАО «Казахский агротехнический исследовательский университет имени Сакена Сейфуллина», Астана, Казахстан
E-mail: a.ismailova@mail.ru; ORCID ID: https://orcid.org/0000-0002-8958-1846.

Аннотация. С развитием технологии секвенирования генома использование вычислительных технологий для прогнозирования функции белков стало одной из важных задач биоинформатики. Ранние исследования в этой области основывались на сходстве последовательностей и предполагали, что белки со схожими аминокислотными последовательностями имеют схожие функции. Однако предложенные ранее методы прогнозирования функций часто не могли выявлять скрытые закономерности между белками и терминами генной онтологии, что понижало точность функционального аннотирования. Глубинное машинное обучение, как показывает множество исследований, справляется с этой задачей на более высоком уровне. Во-первых, методы глубинного обучения могут обучаться на больших объемах данных белковых последовательностей, не принимая во внимание дополнительную информацию о свойствах белков. Во-вторых, подходы глубинного обучения решают такие побочные задачи как зашумлённость данных, их избыточность и высокая размерность. Комбинирование механизма самовнимания и двунаправленной сети с долговременной краткосрочной памятью может быть использовано для решения проблемы функционального

аннотирования белка. Двунаправленная LSTM используется для получения как глобальной, так и локальной информации о свойствах белковых последовательностей, а также для сохранения полученной информации. Алгоритм самовнимания применяется для оптимального использования взаимосвязи последовательности и информации о функциях различных позиций последовательности, что повысит надежность прогнозирования. В качестве инструмента для реализации алгоритмов был выбран язык python, модель обучена в течение 50 эпох и протестирована на экспериментальном наборе данных белка Indica, полученного из открытых источников. Результаты эксперимента показывают, что алгоритм комбинирования механизма самовнимания и двунаправленной сети с долговременной краткосрочной памятью превосходит другие традиционные алгоритмы нейронных сетей и может более точно прогнозировать функцию белка, что показывает возможную применимость алгоритма в функциональном аннотировании белковых последовательностей.

Ключевые слова: двунаправленная LSTM, самовнимание, предсказание функций, белки, машинное обучение

Введение

Предсказание функций белка является серьезной проблемой в области биоинформатики. С развитием технологии секвенирования основным методом прогнозирования функций белка являлся метод биологических экспериментов, который требовал большого количества материальных ресурсов и времени. Увеличивающаяся скорость роста данных о последовательностях белка сделала ручную аннотацию неконкурентоспособной (Саидния и др., 2015), а вычислительные методы аннотации стали основными в области прогнозирования функций белка (Цзян и др., 2017).

Ранние вычислительные методы использовали BLAST, PSI-BLAST, FASTA и другое программное обеспечение для поиска похожих последовательностей каждого белка в обучающем наборе, а затем предполагали, что аналогичные последовательности имеют схожие функции (Гиллис и др., 2013), и мигрировали аннотации функций белков. С развитием искусственного интеллекта многие методы машинного обучения стали широко применяться для прогнозирования белковых функций. Например, SVM-Prot (Цай и др., 2003) использовал состав и трансформацию белка, особенности распределения и алгоритм SVM для прогнозирования функций белка. ProMK (Ю и др., 2015) объединил алгоритм KN с пятью различными методами измерения расстояний между характеристическими значениями для прогнозирования функций белка в различных наборах данных. Многие другие исследователи использовали различные методы машинного обучения для предсказания функций белка и добились хороших результатов, таких как совместное обучение (Нам и др., 2005), наивная байесовская модель (Юсеф и др., 2008), случайный лес (Чен и др., 2005) и другие.

Однако неглубокие методы прогнозирования функций белков часто затрудняют выявление глубоких (нелинейных) взаимосвязей между белками и функциональными терминами Gene Ontology (GO). По сравнению с традиционными методами машинного обучения, методы глубокого обучения могут обучаться на массивных данных о последовательностях белков без разработки признаков. Пока данные аминокислотной последовательности просто обрабатываются, их можно напрямую вводить в нейронную сеть для обучения. Методы глубокого обучения решают проблемы, которые трудно было решить с помощью традиционных алгоритмов машинного обучения в прошлом, такие как высокая размерность, избыточность и высокий уровень шума, вызванные массивными данными о последовательностях белков.

DeepGO (Кулманов и др., 2018) как одна из первых моделей глубокого обучения использовала алгоритм сверточной нейронной сети (CNN) для прогнозирования функции белка с использованием различных наборов данных. Это был алгоритм предсказания функций белков по последовательностям phstein и сетям PPI. На основе алгоритма DeepGO был разработан DeepGOPlus (Кулманов и др., 2019) для прогнозирования функции белка только по аминокислотным последовательностям, в котором модель CNN сочетается с методом BLAST, основанным на сходстве. Он объединил прогнозы нейронной сети с методами, основанными на сходстве последовательностей, для сбора информации о взаимодействии.

В ProtConv (Сара и др., 2021) алгоритм CNN был представлен и обучен для задачи прогнозирования функции белка. Он преобразовал векторное представление последовательности белка или пептида в двумерное изображение с одним каналом, которое подается в CNN.

Несмотря на то, что предложенные выше модели обеспечивают относительно хорошие результаты прогнозирования при решении задачи прогнозирования функции белка, все же существуют некоторые проблемы. С одной стороны, сетевая структура не может эффективно фиксировать долгосрочную зависимость между одной и той же последовательностью белка и не может полностью извлекать информацию о последовательности аминокислот. Долгосрочная зависимость относится к отношениям зависимости на большом расстоянии между каждой аминокислотой в последовательности белка. Установив эту взаимосвязь, можно лучше усвоить общую информацию о последовательности. С другой стороны, трудно эффективно отличить достоверную информацию от недействительной информации о последовательности белка. Трудно уловить аминокислотную последовательность, которая оказывает большее влияние на функцию белка. Достоверная информация относится к информации о последовательности белка, которая оказывает большое влияние на функцию белка. Соответственно, недействительная информация относится к информации о последовательности белка, которая оказывает меньшее влияние на функцию белка.

Метод комбинации механизма самовнимания и BiLSTM может быть

использован для решения проблемы прогнозирования функции белка. Во-первых, для информации о последовательности аминокислот, которая не может быть полностью извлечена, двунаправленная сеть долговременной кратковременной памяти (BiLSTM) (Грейвс и др., 2005) используется для извлечения глобальной и локальной информации о свойствах белков. В то же время связь последовательности между информацией об объектах может быть эффективно сохранена, так что модель может получить лучший эффект прогнозирования. Во-вторых, чтобы лучше использовать взаимосвязь последовательности между информацией о функциях и отражать важность различных позиций последовательности, в этом эксперименте используется механизм самовнимания (Ченг и др., 2016), чтобы заставить модель уделять больше внимания важным функциям в последовательности, тем самым повышая надежность и универсальность модели предсказания функции белка.

Таким образом, в качестве объекта исследования в данной статье используются различные белки, последовательности которых находятся в открытом доступе. Экспериментальные результаты показывают, что, по сравнению с классическим алгоритмом нейронной сети (алгоритм CNN и алгоритм LSTM) и комбинированной версией алгоритма CNN и алгоритма BiLSTM (алгоритм CNN-BiLSTM) алгоритм самовнимания-BiLSTM, использованный в этой статье, позволяет достичь лучших результатов прогнозирования при прогнозировании функции белка.

Материалы и методы

Чтобы использовать нейронную сеть для прогнозирования функции белка на основе аминокислотных последовательностей, первая задача состоит в том, чтобы найти лучший способ представления входных данных, чтобы последовательности белков могли быть распознаны программой. Популярные на сегодняшний день методы кодирования включают метод «горячего» кодирования, «изученные вложения» и «вложения BLOSUM62». По сравнению с методом «изученные вложения» метод «горячего» кодирования позволяет не только уменьшить количество параметров модели, но и избежать проблемы переобучения. Метод BLOSUM62 является одним из популярных методов кодирования. Он представляет каждую аминокислоту соответствующей строкой в матрице BLOSUM62. Вместо того, чтобы рассматривать все аминокислоты независимо друг от друга, матрица BLOSUM62 хранит эволюционную информацию о том, какие пары аминокислот легко взаимозаменяемы в ходе эволюции. Исследование показало, что метод «горячего» кодирования обеспечивает меньшую ошибку модели по сравнению с встраиванием BLOSUM62. Поэтому в данном исследовании последовательности кодов аминокислот кодируется методом «горячего» кодирования. Этот метод сопоставляет каждой букве аминокислоты конкретное действительное число от 1 до 20.

Затем каждому терму n-граммы ставится в соответствие вектор, состоящий из всех нулей, кроме единицы в позиции, зарезервированной для этого терма.

Например, действительное число, соответствующее букве D, равно 3, а это значит, что третьей позиции ее вектора присваивается единица, а остальные позиции равны нулю.

Стоит отметить, что длины белковых последовательностей в основном неодинаковы и сильно варьируют. Чтобы унифицировать формат входных данных и сократить время расчета модели, в этом эксперименте длина каждой белковой последовательности унифицирована до 1002. Несмотря на ограничение, состоящее в том, что длина последовательности составляет 1002 и она не содержит неоднозначных кодов аминокислот, около 90% белковых последовательностей в UniProt удовлетворяют этим условиям (Кулманов и др., 2019).

Другими словами, белковые последовательности длиной более 1002 отфильтровываются. Если последовательности белков с начальной длиной меньше 1002, они дополняются нулями слева до тех пор, пока длина последовательности не станет равной 1002. Наконец, все последовательности белков с неоднозначными кодами аминокислот (B, J, O, U, X, Z) удаляются.

Пример последовательности:

PESRIRLSTRRDAHGMPIPRIESRLGPDAFARLRFMARTCRAILAAAGCAAP FEEFSSADAFSSTHVFGTCRMGHDPMRNVVDGWGRSHRWPNLFVADAS LFPSSGGGESPGLTIQALALRT

Для каждой последовательности один буквенный код заменяется на число. После кодирования вышеупомянутая последовательность будет иметь такой вид:

[13, 7, 13, 4, 16, 15, 8, 15, 10, 16, 17, 15, 15, 3, 1, 7, 6, 11, 13, 8, 13, 15, 8, 4, 16, 15, 10, 6, 13, 3, 1, 5, 1, 15, 10, 15, 5, 11, 1, 15, 17, 2, 15, 1, 8, 10, 1, 1, 1, 6, 2, 1, 1, 13, 5, 4, 4, 5, 16, 16, 1, 3, 1, 5, 16, 16, 17, 7, 18, 5, 6, 17, 2, 15, 11, 6, 7, 3, 13, 11, 15, 12, 18, 18, 3, 6, 19, 6, 15, 16, 7, 15, 19, 13, 12, 10, 5, 18, 1, 3, 1, 16, 10, 5, 13, 16, 16, 6, 6, 6, 4, 16, 13, 6, 10, 17, 8, 14, 1, 10, 1, 10, 15, 17]

В используемом методе, во-первых, для информации о последовательности аминокислот, которая не может быть полностью извлечена, используется двунаправленная сеть долговременной кратковременной памяти (BiLSTM) для извлечения глобальной и локальной информации о свойствах белков (Грейвс и др., 2005). В то же время связь последовательности между информацией об объектах может быть эффективно сохранена, так что модель может получить лучший эффект прогнозирования.

Во-вторых, чтобы лучше использовать взаимосвязь последовательности между информацией об особенностях и отражать важность различных положений последовательности, в этом эксперименте используется механизм self-attention, чтобы заставить модель уделять больше внимания важным функциям в последовательности, тем самым повышая эффективность, надежность и универсальность алгоритма предсказания функции белка.

BiLSTM — это один из типов рекуррентных нейронных сетей, который обрабатывает данные последовательности как в прямом, так и в обратном

направлении с двумя отдельными скрытыми слоями. BiLSTM основан на вентилях ввода, забывания и вывода. Для расчета прогнозных значений используются следующие формулы (1) (Абдулджаббар и др., 2021):

$$\begin{aligned} \text{input gate}(i_t) &= \sigma_g(W_i X_t + R_i h_{t-1} + b_i), \\ \text{forget gate}(f_t) &= \sigma_g\left(W_f X_t + R_f h_{t-1} + b_f\right), \\ \text{cell candidate}(c_t) &= \sigma_g(W_c X_t + R_c h_{t-1} + b_c), \\ \text{output gate}(o_t) &= \sigma_g(W_o X_t + R_o h_{t-1} + b_o), \end{aligned} \quad (1)$$

где σ_g — функция активации вентиля, а W_i, W_f, W_c, и W_o входные весовые матрицы, тогда как R_i, R_f, R_c, и R_o — весовые матрицы, соединяющие предыдущее выходное состояние ячейки с тремя вентилями и входное состояние ячейки. X_t — вход, и h_{t-1} выход в предыдущий момент времени (t–1). b_i, b_f, b_o и b_c — векторы смещения.

На каждой временной итерации t состояние выхода ячейки C_t, и выход слоя h_t можно рассчитать следующим образом (Куртукова и др., 2019):

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t, \quad (2)$$

$$h_t = o_t * \tanh(C_t) \quad (3)$$

Архитектура двунаправленной модели LSTM представлена на Рисунке 1.

Использование двунаправленного LSTM запускает ввод двумя способами, позволяя сохранять контекстную информацию из прошлого и будущего в любой момент времени. Алгоритм BiLSTM может собирать важную информацию об аминокислотных последовательностях в двух направлениях, полностью учитывать информацию о контекстуальной корреляции текущих аминокислотных последовательностей и может более глубоко изучать особенности белковых последовательностей.

Однако из-за длинной аминокислотной последовательности модель BiLSTM не может уловить самую прямую связь между вектором признаков и меткой результата. Добавление в модель механизма self-attention (Васвани и др., 2017) может решить эту проблему. Он может взвешивать входные функции и измерять важность каждой функции для экспериментального объекта. Механизм self-attention широко используется в области классификации текстов и изображений машинного перевода и биоинформатики. В этой экспериментальной модели взаимосвязь вычислений в механизме self-attention показана на Рисунке 2.

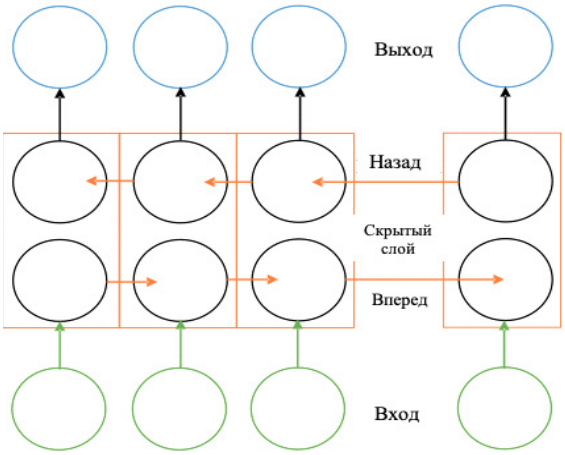

Рисунок 1 – BiLSTM-архитектура

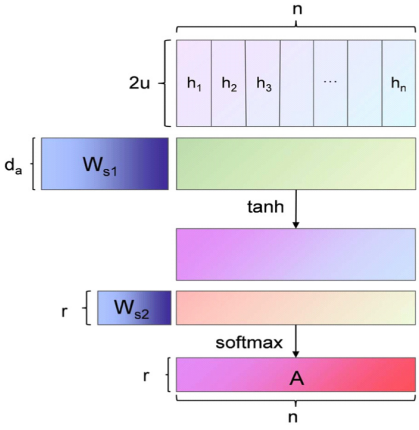

Рисунок 2 - Взаимосвязь вычислений в механизме self-attention

Для признаков, находящихся далеко друг от друга и взаимозависимых, требуется определенное количество времени и шагов, чтобы накопить достаточно информации, чтобы связать их. Чем дальше они друг от друга, тем меньше вероятность того, что сеть BiLSTM захватит эффективную информацию. Это означает, что, когда аминокислота может быть связана с окружающими ее аминокислотами или более отдаленными аминокислотами, использование алгоритма BiLSTM учитывает только информацию до и после последовательности белка в определенном диапазоне и не может решить проблему корреляции между прерывистыми аминокислотами. Стоит отметить, что одна аминокислота или несколько аминокислот могут иметь большое влияние на функцию белка. В процессе расчета механизм self-attention может напрямую связать корреляцию между любыми двумя функциями в последовательности за один шаг расчета, что значительно

сокращает расстояние между зависимыми функциями на большом расстоянии. Следовательно, комбинация алгоритма BiLSTM и self-attention уделяет больше внимания аминокислотам, которые могут иметь большое влияние на функцию белка, так что аминокислотная последовательность вносит больший вклад в точное предсказание функции белка.

Техническая реализация были исполнена на языке Python с использованием библиотек Num.py, TensorFlow и Keras и запущена на облачной платформе Colab (https://colab.research.google.com/).

При реализации модели были использованы следующие инструменты:

− В качестве алгоритма оптимизации был выбран алгоритм Adam;

− показатель точности Accuracy как целевая функция;

− Бинарно-кроссэнтропийная функция, возвращающая ошибку классификации как функцию логистических потерь Loss:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i * \log(\hat{y}_i) + (1 - \hat{y}_i) * \log(1 - \hat{y}_i), \quad (4)$$

где y_i – истинная метка класса; $\hat{y}_i$ ответ классификатора (вычисляемая метка класса) на *i*-й объект; *N* – количество классов.

Эффективность алгоритмов прогнозирования функции белка оценивается по четырем популярным показателям (Cheng., 2016): чувствительности (sensitivity - SE), специфичности (specificity – SP), точности (accuracy - ACC) и коэффициенту корреляции Мэтьюса (Matthews correlation coefficient – MCC). Все четыре показателя широко используются для оценки эффективности предикторов функции белков и основываются на четырех составляющих *TP*, *TN*, *FP* и *FN*, представляющих собой истинно положительные, истинно отрицательные, ложноположительные и ложноотрицательные результаты соответственно.

В частности, SE определяется процентом истинно положительных образцов, правильно идентифицированных как «положительные»:

$$SE = \frac{TP}{TP+FN} \quad (5)$$

SP указывает долю истинно отрицательных образцов, которые были правильно предсказаны как «отрицательные»:

$$SP = \frac{TN}{TN+FP} \quad (6)$$

ACC относится к количеству истинных образцов (положительных плюс отрицательных), деленному на количество всех изученных образцов:

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

MCC является важным показателем, отражающим стабильность предиктора функции белка, который описывает корреляцию между прогностическим значением и фактическим значением. Он считается одним из наиболее полных параметров в любой категории предикторов из-за полного учета всех четырех результатов. В частности, MCC (F1) можно рассчитать по следующей формуле:

$$MCC = \frac{(TP*TN-FP*FN)}{\sqrt{(TP+FN)*(TP+FP)*(TN+FP)*(TN+FN)}} \quad (8)$$

Результаты и обсуждение

Для обучения модели были запущены 50 эпох обучения. Для предотвращения переобучения модели были выставлены следующие параметры: earlystopping = EarlyStopping(monitor='val_loss', patience=3, verbose=1). В Таблице 1 представлены результаты обучения сети, полученные в течение первых 39 эпох обучения.

Таблица 1. Результаты обучения сетевой модели

Epoch	Loss	Accuracy	Val_loss	Val_accuracy
1	5.4675	0.1226	3.4304	0.3755
2	2.8003	0.4818	1.9909	0.6735
3	1.8967	0.6634	1.3769	0.7867
4	1.4488	0.7507	1.0554	0.8479
5	1.2040	0.7947	0.8728	0.8772
6	1.0520	0.8230	0.7566	0.8987
7	0.9574	0.8404	0.7134	0.9029
…				
33	0.5505	0.9144	0.3783	0.9610
34	0.5439	0.9147	0.3818	0.9609
35	0.5398	0.9163	0.3673	0.9635
36	0.5344	0.9167	0.3582	0.9648
37	0.5306	0.9174	0.3720	0.9616
38	0.5259	0.9186	0.3598	0.9638
39	0.5209	0.9189	0.3611	0.9639

Оценка производительности модели показала, что наилучшая точность BiLSTM составила 0.9699 для обучающего набора, 0. 9638 для набора данных и 0.9632 для тестового набора.

На Рисунке 3 показаны значения всех четырёх показателей для разработанной модели на примере тестового набора данных белк5a Indica, также приведены результаты запуска нескольких классических моделей на тех же наборах данных.

Экспериментальные результаты алгоритма SA-BiLSTM сравниваются с алгоритмами CNN, LSTM и CNN-BLSTM.

Набор данных	Подтип онтологии	Алгоритм	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Indica	BP	CNN	73.856	50.237	88.228	64.020
		LSTM	89.140	81.669	77.071	79.294
		CNN-BiLSTM	87.973	78.994	74.139	76.461
		SA-BiLSTM	90.551	83.113	79.443	81.186
	MF	CNN	84.454	73.636	51.654	60.673
		LSTM	86.190	66.831	61.324	63.911
		CNN-BiLSTM	84.796	63.966	63.482	63.660
		SA-BiLSTM	87.274	75.071	63.248	68.608
	CC	CNN	81.506	78.081	77.655	77.800
		LSTM	82.186	80.101	77.618	78.825
		CNN-BiLSTM	81.691	78.400	77.655	78.006
		SA-BiLSTM	85.299	81.409	83.919	82.434

Рисунок 3 - Результаты предсказания функций белков на тестовых данных

Набор данных	Подтип онтологии	Белок	Функция	Предсказанная функция
Indica	BP	Q01N44	GO:0016042	GO:0016042
			GO:0006629	GO:0006629
				GO:0006807
	MF	E0ZS48	GO:0009039	GO:0009039
			GO:0016787	GO:0016787
			GO:0046872	GO:0046872
			GO:0016810	
	CC	Q01N44	GO:0005783	GO:0005783
			GO:0005886	GO:0005886
			GO:0016020	GO:0016020
			GO:0005789	GO:0005789
				GO:0009536

Рисунок 4 - примеры предсказания функций белков Indica

Белок E0ZS48 (UREA_ORYSI), как было обнаружено вручную, имеет функции белка GO: 0009039, GO: 0016787, GO: 0016810 и GO: 0046872. Среди них функция GO:0009039 представляет активность уреазы. Функция GO:0016787 указывает на гидролазную активность, которая может катализировать гидролиз различных связей. Функция GO:0016810 означает активность гидролазы, которая катализирует гидролиз любой углерод-азотной связи C-N, кроме пептидных связей. GO:0046872 означает, что белок выполняет функцию связывания с ионом металла. К сожалению, экспериментальные результаты показывают, что функция GO:0016810 не может быть успешно предсказана, что может быть связано с тем, что функция GO:0016910 и функция GO:0016787 имеют схожие функции. Обе функции катализируют гидролиз определенных связей. Поэтому для экспериментов трудно сделать абсолютно точное предсказание функции по этой проблеме.

Заключение

Таким образом, предложенная модель не обязательно точно предсказывает полную функцию белка с очень похожими функциями. Но функции большинства белков можно точно предсказать. Согласно Рисунку 3 и Рисунку 4, эффект предсказания белка с помощью алгоритма SA-BiLSTM в целом

хороший. Более того, в дополнение к полному предсказанию аннотации функции белка эксперимент также может предсказать аннотации функции GO, которые не показаны в базе данных Swiss-Prot. Это дает новое направление для последующих экспериментальных исследований.

ЛИТЕРАТУРА

Абдулджаббар Р.Л., Диа Х., Цай П. Однонаправленные и двунаправленные LSTM-модели для краткосрочного прогнозирования трафика // Journal of Advanced Transportation. – 2021. – Том 4. – С. 1-16.

Грейвс А., Шмидхубер Дж . Покадровая классификация фонем с использованием двунаправленного LSTM и других архитектур нейронных сетей // Нейронные сети. – 2005. – С. 602-610.

Гиллис Дж., Павлидис П. Характеристика уровня техники в вычислительном присвоении функции гена: уроки первой критической оценки функциональной аннотации (CAFA) // BMC Bioinformatics. – 2013.

Кулманов М., Хендорф Р., DeepGO: прогнозирование функций белка по последовательности и взаимодействиям с использованием классификатора, основанного на глубокой онтологии // Биоинформатика. – 2018. – Том 34. – С. 660-668.

Кулманов М., Хендорф Р. DeepGOPlus: Улучшенное предсказание функции белка по последовательности // Биоинформатика. – 2019.

Куртукова А.В., Романов А.С. Моделирование архитектуры нейронной сети для идентификации автора исходного кода // Известия Томского государственного университета систем управления и радиоэлектроники. – 2019. – Том 22. – С. 37-42.

Сара С., Хасан М., Ахмад А., Шатабда С. Сверточные нейронные сети с графическим представлением аминокислотных последовательностей для прогнозирования функций белков // Вычислительная биология и химия. – 2021.

Ченг Дж., Донг Л., Лапата М. Долговременная кратковременная память-сети для машинного чтения // Конференция по эмпирическим методам обработки естественного языка. – 2016. – С. 551-561.

Цай С.З., Хан Л.Ю., Цзи З.Л., Чен Х., Чен Ю.З. SVM-Prot: веб-программное обеспечение для обработки опорных векторов для функциональной классификации белка по его первичной последовательности. Исследование нуклеиновых кислот. – 2003. – С. 3692-3697.

Чен Х.В., Лю М. Прогнозирование белок-белковых взаимодействий с использованием структуры леса случайных решений // Биоинформатика. – 2005. – С. 4394-4400.

Саидния С., Манайи А., Абдоллахи М. От экспериментов in vitro к in vivo и клиническим исследованиям; плюсы и минусы // Современные технологии разработки лекарственных средств. – 2015. – С. 218-224.

Нам Дж.У., Шин К.Р., Хан Дж., Ли Ю., Ким В.Н., Чжан Б.Т. Предсказание микроРНК человека с помощью вероятностной модели совместного обучения последовательности и структуры // Исследования нуклеиновых кислот. - 2005. – С. 3570-3581.

Васвани А., Шазир Н., Пармар Н. и др. Внимание - это все, что вам нужно // NIPS'17: Материалы 31-й международной конференции по нейронным системам обработки информации. – 2017. – С. 6000-6010.

Цзян Ю., Орон Т., Кларк У. и др. Расширенная оценка методов прогнозирования функций белков показывает повышение точности // Биология генома. – 2017.

Ю. Г., Рангвала Х., Доменикони С., Чжан З. Чжан З. Предсказание функции белка с использованием нескольких ядер // Транзакции IEEE/ACM по вычислительной биологии и биоинформатике. – 2015. – С. 219-233.

Юсеф М., Юнг С., Шоу Л.С. и др. Обучение на положительных примерах, когда класс отрицательных не определен - идентификация генов микроРНК // Алгоритмы молекулярной биологии. – 2008.

REFERENCES

Abduljabbar R.L., Dia H., Tsai P. Unidirectional and Bidirectional LSTM Models for Short-Term Traffic Prediction // Journal of Advanced Transportation. – 2021. – Vol. 4. – P. 1-16.

Graves A., Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures // Neural Networks. – 2005. – P. 602-610.

Gillis J., Pavlidis P. Characterizing the state of the art in the computational assignment of gene function: lessons from the first critical assessment of functional annotation (CAFA) // BMC Bioinformatics. – 2013.

Kulmanov M., Hoehndorf R, DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier // Bioinformatics. – 2018. – Vol. 34. – P. 660–668.

Kulmanov M., Hoehndorf R. DeepGOPlus: Improved protein function prediction from sequence // Bioinformatics. – 2019.

Kurtukova A.V., Romanov A.S. Modeling the neural network architecture to identify the author of the source code // Proceedings of Tomsk State University of Control Systems and Radioelectronics. – 2019. – Vol. 22. – P. 37-42.

Sara S., Hasan M., Ahmad A., Shatabda S. Convolutional neural networks with image representation of amino acid sequences for protein function prediction // Computational Biology and Chemistry. – 2021.

Cheng J., Dong L., Lapata M. Long Short-Term Memory-Networks for Machine Reading // Conference on Empirical Methods in Natural Language Processing. – 2016. – P. 551-561.

Cai C.Z., Han L.Y., Ji Z.L., Chen X., Chen Y.Z. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Research. – 2003. – P. 3692-3697.

Chen X.W., Liu M. Prediction of protein-protein interactions using random decision forest framework // Bioinformatics. – 2005. – P. 4394-4400.

Saeidnia S., Manayi A., Abdollahi M. From in vitro Experiments to in vivo and Clinical Studies; Pros and Cons // Current Drug Discovery Technologies. – 2015. – P. 218-224.

Nam J.W., Shin K.R., Han J., Lee Y., Kim V.N., Zhang B.T. Human microRNA prediction through a probabilistic co-learning model of sequence and structure // Nucleic Acids Research. – 2005. – P. 3570-3581.

Vaswani A., Shazeer N., Parmar N., et al. Attention is all you need // NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. – 2017. – P. 6000-6010.

Jiang Y., Oron T., Clark W. et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy // Genome Biology. – 2017.

Yu G., Rangwala H., Domeniconi C., Zhang G., Zhang Z. Predicting Protein Function Using Multiple Kernels // IEEE/ACM Transactions on Computational Biology and Bioinformatics. – 2015. – P. 219-233.

Yousef M., Jung S., Showe L.C. et al. Learning from positive examples when the negative class is undetermined- microRNA gene identification // Algorithms for Molecular Biology. – 2008.

МАЗМҰНЫ

СОДЕРЖАНИЕ

CONTENTS

Publication Ethics and Publication Malpracticein
the journals of the National Academy of Sciences of the Republic of Kazakhstan

For information on Ethics in publishing and Ethicalguidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the National Academy of Sciences of the Republic of Kazakhstan implies that the described work has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see http://www.elsevier.com/postingpolicy), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any otherlanguage, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The National Academy of Sciences of the Republic of Kazakhstan follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (http://publicationethics.org/files/u2/New_Code.pdf). To verify originality, your article may be checked by the Cross Check originality detection service http://www.elsevier.com/editors/plagdetect.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the National Academy of Sciences of the Republic of Kazakhstan.

The Editorial Board of the National Academy of Sciences of the Republic of Kazakhstan will monitor and safeguard publishing ethics.

Правила оформления статьи для публикации в журнале смотреть на сайтах:
www:nauka-nanrk.kz
http://physics-mathematics.kz/index.php/en/archive
ISSN2518-1726 (Online),
ISSN 1991-346X (Print)

Подписано в печать 28.09.2023.
Формат 60x881/8. Бумага офсетная. Печать –ризограф.
18,0 п.л. Тираж 300. Заказ 3.

РОО «Национальная академия наук РК» 050010,
Алматы, ул. Шевченко, 28, т. 272-13-19