8D06101 – «Үлкен деректер аналитикасы» білім беру бағдарламасы бойынша философия докторы (PhD) дәрежесін алу үшін Голенко Екатерина Сергеевнаның «Нативті ақуыздардың масс-спектрометрия деректерін талдау алгоритмдерін жасау» тақырыбындағы диссертациялық жұмысына

АНДАТПА

Зерттеу тақырыбының мәселесі және өзектілігі. Қазіргі ғылыми тәжірибеде масс-спектрометрия өзін пептидтер мен ақуыздарды талдаудың орталық технологияларының бірі ретінде көрсетті. Масс-спектрометрия көмегімен ақуыздарды идентификациялау процедурасы белоктарды пептидтерге ыдыратуды қамтиды, содан кейін олар бөлінеді, фрагменттеледі, иондалады және масс-спектрометрлер арқылы сүзіліп алынады. Белоктар олардың масс-спектрлеріне, дәлірек айтқанда, пептидтік фрагмент иондарына сәйкес келетін сипаттамалық шыңдарына қарай танылады және каталогталады. Көптеген факторлар, соның ішінде трансляциядан кейінгі модификациялардың болуы, ақуыздың фрагментациясы және иондану проблемалары белоктардың шектеулі жиынтығын ғана дәл идентификациялауға болатындығын білдіреді. Заманауй әдістер, әдетте, үлгідегі барлық ақуыздардың жартысынан азын сенімді түрде идентификациялайды, олардың болуы мен функциясын растаусыз ықтимал маңызды молекулалардың айтарлықтай санын қалдырады. Бұл күрделі биологиялық жүйелерді түсінуге айтарлықтай кедергі жасайды және масс-спектрометрия әдістерін одан әрі жетілдіруді және спектрлік деректерді өңдеу мен интерпретациялаудың неғұрлым қуатты алгоритмдерін әзірлеуді талап етеді.

Биоинформатикадағы ақуызды анықтаудың қазіргі стратегиялары екі негізгі санатқа бөлінеді. Біріншісі – белгілі және реттелген белок тізбегі негізінде *in silico*-да жасалған теориялық генерацияланған пептидтік спектрлер пулымен эксперименталды түрде алынған масс-спектрлерді салыстыру принципі бойынша жұмыс істейтін мәліметтер базасын іздеу. Протеин дерекқорлары әдетте Mascot, SEQUEST және X!Tandem сияқты арнайы іздеу жүйелері арқылы қолданылады. Спектрлік кітапхана арқылы іздеу протеомикадағы жетілдірілген әдіснаманы білдіреді. Бұл әдіс алдыңғы эксперименттерден алынған пептидтік масс-спектрлер деректерін қамтитын кітапханаларды пайдаланады. Белгісіз пептидтің массалық спектрі пайда болған кезде дәл сәйкестендіру үшін ол кітапханадағы бар спектрлермен салыстырылады. Бұл талдау жылдамдығын жақсартып қана қоймайды, сонымен қатар пептидтерді анықтаудың сезімталдығы мен дәлдігін жақсартады, бұл оны заманауи протеомикада перспективалы тәсілге айналдырады.Соңғы санат - белок реттілігінің дерекқорын пайдаланбай, фрагменттелген пептид иондарына сәйкес шыңдарды зерттеу арқылы масс-спектрометрия деректерінен тікелей ақуыз ретін анықтауға мүмкіндік беретін *de novo* секвенирлеу әдісі.

Пептидтер мен белоктарды анықтау үшін деректер қорын пайдалану - тандемдік масс-спектрометрия деректерін талдау үшін қолайлы және кеңінен қолданылатын әдіс. Бұл әдіс реті дерекқорларда қамтылған зерттелген

белоктарды тану үшін тиімді. Дегенмен, ол белгісіз белоктарды немесе трансляциядан кейінгі өзгерістерге ұшырағандарды іздеу кезінде шектелген, бұл іздеу уақытын және дұрыс емес нәтижелерді алу қаупін арттыруы мүмкін.

Белгілі ақуыздың атқаратын функцияларын анықтауға бағытталған биоинформатикадағы тағы бір маңызды мәселе - ақуыз функциясын болжау. Әдетте, ақуыз функциясын анықтау қолмен немесе компьютерлік аннотация арқылы жүзеге асырылады. Gene Ontology жүйесіне негізделген функцияны автоматтандырылған болжау биоинформатикадағы күрделі міндет болып табылады. Бұл тапсырманың күрделілігі келесі факторларға байланысты: біріншіден, эксперттер аннотацияланбаған ақуыздардың көпшілігінде аминқышқылдарының тізбегінен басқа ешқандай ақпарат жоқ. Екіншіден, ақуыз реттілігі туралы деректер мен әртүрлі биологиялық дерекқорлардағы қосымша ақпарат әртүрлі форматтарда сақталуы мүмкін, бұл машиналық оқыту модельдері үшін кіріс деректер жиынын дайындауда қосымша қиындықтар тудырады. Үшіншіден, Gene Ontology күрделі және біркелкі емес құрылымға ие, сондықтан мүмкіндіктерді болжау мәселесін көп шығыс таңбалау мәселесі ретінде қарастыру керек, бұл мәселені шешуді күрделі етеді.

Осылайша, бүгінгі күні генетикалық және белоктық тізбектерді өңдеу саласында келесі алгоритмдер маңызды болып табылады: (i) бастапқы ақпаратты өңдеуге бағытталған, бастапқы спектрометриялық деректерді тереңірек талдау үшін қолайлы форматқа түрлендіруді жеңілдететін деректерді талдау алгоритмдері; (ii) масс-спектрометриялық кластерлердегі жалпы және жасырын үлгілерді анықтау үшін спектрлік деректерді интерпретациялауға арналған алгоритмдер; (iii) пептидтер мен ақуыздар тізбегін анықтауға арналған дерекқорлар мен байланысты алгоритмдер және (iv) белок тізбегін биологиялық функциялармен байланыстыруға қабілетті функционалдық аннотация алгоритмдері.

Диссертациялық зерттеудің объектісі – АБҒЗП, PhD докторы В.С.Киянның жетекшілігімен «Описторхоз және меторхоз қоздырғыштарын анықтау және дифференциалды диагностикалау үшін ПТР тесті» жобасы аясында алынған масс-спектрометриялық деректер, сондай-ақ жалпыға қолжетімді деректер NIST, Pfam және UniProt қоса алғанда, ақуыз және ДНҚ дерекқорларының реттілігі.

Диссертацияның мақсаты масс-спектрометрия нәтижелерін интерпретациялау және ақуыз функцияларын болжау алгоритмдерін жасау болып табылады.

Диссертациялық зерттеудің міндеттері:

1. Масс-спектрометрия мәліметтері мен ақуыздар реттілігін өңдеуге арналған қолданыстағы шешімдерге талдау жүргізу;
2. Машиналық оқыту модельдерін пайдалана отырып, пептидтерді идентификациялау алгоритмін жасау;
3. Машиналық оқыту әдістерін пайдалана отырып, белок тізбегінің функцияларын болжау алгоритмін жасау;

4. Машиналық оқыту үшін жалпы қабылданған бағалау көрсеткіштерін пайдаланып, жалпыға қолжетімді деректер жиындарында ұсынылған алгоритмдерді бағалау.

Диссертациялық зерттеу пәні пептидтер мен белоктарды анықтау, сонымен қатар олардың функцияларын анықтау алгоритмдері болып табылады.

Зерттеу әдістері. Эксперименттік деректерді талдау және алгоритмдерді құру кезінде үлкен деректер жиынын талдау әдістері, сапалық талдау, реттілікті салыстыру әдістері, нейрондық желілер, кластерлеу және классификациялау әдістері қолданылды.

Қорғауға ұсынылатын және ғылыми жаңалық белгілері бар негізгі ғылыми ережелер:

1. Спектрлер және пептидтермен жұмыс істеу үшін терең ұқсастық желісіне енгізілген екі бағытты LSTM нейрондық желісіне негізделген масс-спектрометрия арқылы алынған пептидтерді анықтау алгоритмі;

2. Екі бағытты LSTM нейрондық желісі және «self-attention» механизмі негізінде белок тізбегі функцияларын болжау алгоритмі.

Тақырыптың ғылыми-зерттеу бағдарламаларының жоспарларымен байланысы. Зерттеу тақырыбы «Ақпараттық, коммуникациялық және ғарыштық технологиялар» басым ғылыми бағытына сәйкес келеді.

Ұсынылған нәтижелер 2018-2020 жылдары AP05131132 «Описторхоз және меторхоз қоздырғыштарын анықтау және дифференциалды диагностикалау үшін ПТР-тесті» жобасын орындау кезінде алынды.

Диссертациялық зерттеудің ғылыми жаңалығы:

- Пептидтер спектрінің сәйкестіктерін іздеу үшін BiLSTM нейрондық желісін пайдалана отырып, SpeCollate ашық бастапқы ұқсастық желісі негізінде әзірленген пептидтерді идентификациялау алгоритмі ұсынылды.
- BiLSTM нейрондық желісі мен өзіне-өзі назар аудару (self-attention) механизмінің комбинациясына негізделген белок функцияларын аннотациялау алгоритмі ұсынылған.
- Әзірленген алгоритмдер бір уақытта машиналық оқытуды да, реттілікті салыстыру әдістерін де қолдана отырып, биологиялық деректерді өңдеуге мүмкіндік береді.

Диссертациялық зерттеу нәтижелерінің теориялық маңызы:

- Әзірленген алгоритмдерде аминқышқылдарының тізбегінің ұзындығына шектеулер жоқ, сондықтан геном масштабында белок функцияларына аннотация жасау үшін пайдалануға болады.
- Олар жылдам жұмыс істейді және бірнеше минут ішінде, тіпті бір процессорда бірнеше мың ақуызға түсініктеме бере алады.
- Модельдер ақуызаралық әрекеттесуі туралы теңгерімсіз немесе жетіспейтін ақпаратпен шектелмейді.
- Алгоритмдерді тек реттілік мотивтері арқылы қолдануға болады.

Диссертациялық зерттеу нәтижелерінің практикалық маңызы:

- Әзірленген алгоритмдерді зертханалық бағдарламалық модульдерде ақуыз ретін анықтау және олардың функцияларын жоғары сенімділікпен болжау үшін енгізуге болады.

- Әзірленген алгоритмдерді биологтар биологиялық деректермен жұмыс істеу кезінде қолданыстағы қолданбаларға немесе қосымша құралдарға балама ретінде пайдалана алады.

- Алгоритмдердің негізінде жатқан нейрондық желілерді, ең алдымен, ауру механизмдерін және дәрілік препараттарды жасау ретін түсіну үшін белоктардың функцияларын анықтауды талап ететін кең ауқымды биологиялық мәселелерді шешу үшін әртүрлі деректер жинақтарында оқытуға болады.

Диссертациялық жұмыстың ғылыми негізделген теориялық және эксперименттік нәтижелері «Описторхоз және меторхоз қоздырғыштарын анықтау және дифференциалды диагностикалау үшін ПТР тесті» тақырыбындағы ғылыми жобада пайдаланылды.

Диссертациялық зерттеу нәтижесінде құрылған бағдарламалық модульдер «Ұлттық биотехнология орталығының» биоәртүрлілік және генетикалық ресурстар зертханасында (Астана қ., Қазақстан) және «Жаңа бағдарламалық жүйелер» ЖШҚ (Новосібір, Ресей Федерациясы) қолданылды және енгізілді.

Диссертациялық зерттеу нәтижелерін апробациялау. Диссертациялық зерттеудің негізгі нәтижелері С.Сейфуллин атындағы ҚАТЗУ «Ақпараттық жүйелер» кафедрасының, Л. Н. Гумилёв атындағы ЕҰУ «Ақпараттық жүйелер» кафедрасының ғылыми семинарларында баяндалып, талқыланды және келесі халықаралық ғылыми-практикалық конференцияларда:

- «Сейфуллин оқулары – 16» Халықаралық ғылыми-теориялық конференциясы, С.Сейфуллин аындағы ҚАТЗУ, Нұр-Сұлтан, 2020 ж.;

- «Ғылым, білім және өндіріс интеграциясы Ұлт жоспарын жүзеге асырудың негізі» Халықаралық ғылыми-тәжірибелік конференциясы, ҚарМТУ, 2020 ж.;

- «Сейфуллин оқулары – 17: «Қазіргі ауыл шаруашылығы ғылымы: цифрлық трансформация» Халықаралық ғылыми-теориялық конференциясы, Нұр-Сұлтан, 2021 ж.;

- «XXII Сәтбаев оқулары» Халықаралық ғылыми конференциясы, Сәтбаев университеті, Алматы, 2022 ж.;

- «Сейфуллин оқулары – 18: «Жастар және ғылым – болашаққа көзқарас» атты халықаралық ғылыми-теориялық конференция», С.Сейфуллин атындағы ҚАТЗУ, Астана, 12 сәуір 2022 жыл;

- «Математикалық логика және информатика» Халықаралық ғылыми конференциясы, Л.Н.Гумилев атындағы ЕҰУ, Астана, 7-8 қазан 2022 ж.

SCOPUS халықаралық базасына кіретін журналдардағы мақалалар. Автордың қосқан үлесі гипотеза жасау, деректерді жинау, эксперименттерді техникалық іске асыру, нәтижелерді түсіндіру және жарияланымды дайындау болып табылады:

Golenko et al. Implementation of machine learning models to determine the appropriate model for protein function prediction. Eastern-European Journal of Enterprise Technologies, 2022. https://doi.org/10.15587/1729-4061.2022.263270

Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Ғылым және жоғары білім саласындағы сапаны қамтамасыз ету Комитеті ұсынған журналдардағы мақалалар. Осы жарияланымдар аясында автор

деректерді талдаудың зерттеу тұжырымдамасы мен әдіснамасын жасаушы болып табылады. Деректерді жинауға, нәтижелерді интерпретациялауда, қорытындыларды тұжырымдауға және ғылыми мақалаларды жариялауға дайындауға тікелей қатысты:

1. Голенко Е., Исмаилова А., Жумаханова А. Предсказание функций белков при помощи базы данных «Gene Ontology» и моделей машинного обучения. ҚР ҰҒА Жаршысы. Физика және математика сериясы. №2 (342). – 2022. – Б. 19–38.
2. Голенко Е., Исмаилова А., Молдашева Р. Применение методов глубокого обучения для предсказания структуры белков. ҚР Ұлттық инженерлік академия Хабаршысы. Ақпараттық-коммуникациялық технологиялар. №4 (86). – 2022. – Б. 28–40.
3. Голенко Е., Исмаилова А. Предсказание функций белка с использованием комбинации BiLSTM и алгоритма самовнимания. ҚР ҰҒА Жаршысы. Физика және математика сериясы. №3 (347). – 2023. – Б. 62–75.

Автордың жеке үлесі диссертацияның барлық тараулары мен логикалық сілтемелері бойынша тікелей зерттеу жүргізуден тұрады: бұрын ұсынылған жұмыстарға шолу мен талдау жүргізу, пайдаланылған әдістерді таңдау және негіздеу, алгоритмдерді әзірлеу және техникалық енгізу, әзірленген үлгілерді бастапқы деректерде апробациялау және тестілеу.

Диссертациялық зерттеу тақырыбы бойынша жарияланымдар. Диссертациялық зерттеу тақырыбы бойынша 13 (он үш) ғылыми жұмыс жарияланды, оның ішінде 1 (бір) SCOPUS халықаралық деректер базасына енгізілген импакт-факторы нөлден аспайтын ғылыми журналда (CiteScore2022 пайыздық көрсеткіш 34-ке тең), Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Ғылым және жоғары білім саласындағы сапаны қамтамасыз ету комитеті ұсынған журналдарда 3 (үш) мақала, Халықаралық конференциялар жинақтарында 6 (алты) мақала, басқа басылымдарда 3 (үш) мақала. Компьютерлік бағдарламаны Мемлекеттік тіркеу туралы 2 (екі) авторлық куәлік алынды.

Диссертациялық жұмыстың құрылымы мен көлемі. Диссертациялық зерттеу келесі форматта ұсынылған: кіріспе, үш негізгі бөлім, қорытынды, пайдаланылған әдебиеттер тізімі (123 атау) және екі қосымша. Жалпы көлемі 23 сурет пен 7 кестенің сүйемелдеуімен иллюстрациялар, диаграммалар және кестелер сияқты негізгі ойларды бөлектеуге арналған құралдарды пайдаланатын компьютерлік мәтіннің 123 бетін құрайды.

Кіріспеде зерттелетін тақырыптың маңыздылығы, зерттелу деңгейі мен ғылыми күрделілігі көрсетіледі, протеомика мәліметтерін өңдеудің әз ірленген алгоритмдерінің өзектілігі ашылады, зерттеу мақсаты тұжырымдалады, мақсатқа жету үшін міндеттер қойылады, пән мен объект анықталады. зерттеудің ғылыми жаңалығын, диссертациялық зерттеудің теориялық және практикалық маңыздылығын ашады. Зерттеу нәтижелерін сынақтан өткізу және жариялау туралы деректер келтіріліп, автордың ғылыми зерттеулерге қосқан жеке үлесі де көрсетіледі.

Бірінші бөлімде белок құрылымдары мен генетикалық тізбектердің халықаралық жаһандық репозиторийлерінің қазіргі жағдайы талданады, осыдан диссертациялық зерттеудің екі негізгі мақсаты тұжырымдалады. Масс-спектрометриялық талдау арқылы бөлінген белоктар мен пептидтерді анықтау мәселесіне кеңінен талдау жасалып, анықталған пептидтердің дұрыстығын бағалау әдістері зерттелді. Ақуыздар мен пептидтерді анықтау мәселесін шешудің ерте және заманауи әдістері қарастырылған. Сонымен қатар, эксперименталды түрде алынған белоктарды аннотациялау мәселелеріне талдау және функционалдық болжау үшін предшественниктер ұсынған алгоритмдерге сыни шолу жүргізілді. Негізгі кемшіліктер анықталды және олардың дәлдігі мен сенімділігін арттыру үшін осы алгоритмдерді жетілдіру бағыттары белгіленді.

Екінші бөлім бір евклид кеңістігінде спектрлер мен пептидтердің бірдей өлшемді кірістірулерін құруға бейімделген, оларды оңтайландыру мен желінің дамуын салыстыруға мүмкіндік беретін SpeCollate жалпыға қолжетімді терең ұқсастық желісі негізінде пептидтер мен ақуыздарды идентификациялау шешімін ұсынады. Осы мақсатқа жету үшін екі бағытты LSTM пайдалануды қоса алғанда, үлгі параметрлерінің нақты жиынтығы таңдалды. Желідегі оқыту процесі де сипатталады, ол оң және теріс мысалдарды көрсететін деректер бойынша оқытудан тұрады және сәйкес және сәйкес емес жұптар арасындағы тиімді дискриминацияға ықпал ететін SNAP жоғалту функциясының контекстінде жүзеге асырылады. Модельді оқыту нәтижелері ұсынылып, оның тиімділігі бағаланады.

Үшінші бөлімде белок тізбегінің функционалдық аннотациясының алгоритмін жасау процесі ұсынылған. Нейрондық желіні оқыту үшін ашық көздерден алынған эксперименттік мәліметтерді алдын ала өңдеу және талдау процесі сипатталған. Екібағытты LSTM қолданатын алгоритмнің математикалық моделі өзіне «self-attention» механизмімен біріктірілген. Эксперименттік мәліметтерді пайдалана отырып, модельді оқыту нәтижелері берілген. Әзірленген модельдің функцияларын болжау сенімділігі машиналық оқыту модельдерінің дәлдігі мен сенімділігін талдау үшін пайдаланылатын жалпы қабылданған параметрлерді пайдалана отырып бағаланды. Ақуыз функцияларын қолмен аннотациялау нәтижелері де әзірленген алгоритмнің жоғары дәлдігін растайды.

Қорытындыда диссертациялық зерттеудің нәтижелері ұсынылып, қорғауға ұсынылған ережелердің ақиқаттығын растайтын және дәлелдейтін негізгі қорытындылар тұжырымдалған.

Қосымшаларда авторлық құқық куәліктері мен енгізу актілері ұсынылған.