Жоба аты: ЖТН AP19678041 «Толық геномдық секвенирлеуде тандемдік қайталануларды идентификациялауға арналған бағдарламалық қамтамасыз етуді әзірлеу»

Өзектілігі:

Соңғы онжылдықтарда геномдағы қайталанатын тізбектердің рөлі туралы идея күрт өзгерді және «қоқыс ДНҚ» санатынан қайталанатын элементтер олардың иелерінің геномдарының жұмыс істеуіне және генетикалық әртүрлілікке және жаңа реттеуші элементтердің пайда болуына ықпал ету арқылы эволюциясына үлкен әсер етеді. Секвенирлеу технологиясының, атап айтқанда, үшінші буынды секвенирлеудің одан әрі дамуы тандемдік қайталауларды зерттеуге айтарлықтай ықпал етеді, бұл егжей-тегжейлі зерттеу үшін жаңа деректердің пайда болуына әкелді. Қысқа тандемді қайталаулар адам геномының шамамен 7% құрайтыны анықталды. Эукариоттар мен прокариоттардың геномдарындағы кең ұсынылуы және олардың өзгергіштігінің жоғары жылдамдығы, геном эволюциясының және генетикалық диверсификациясының негізгі факторларының бірі ретінде қайталанулар олардың рөлі ретінде жүйелі түрде бағаланатын болады. Осы элементтердің көпшілігі ауруларда белсендірілетіні белгілі болғандықтан, тандемдік қайталануларды талдау және биомаркерлік ассоциацияларды анықтау және организмдегі биологиялық процестерді реттеу кезінде генетикалық өзгерістерге және күтілетін салдарға қатысты дербестендірілген медицина мен ауруды диагностикалау мүмкіндігі туындайды. Осыған байланысты тандемді қайталаудың әртүрлі формаларын идентификациялауға арналған жетілдірілген және қолдануға оңай биоақпараттық құралдарды әзірлеу **кезек күттірмейтін, өзекті міндет** болып табылады.

Ұсынылған жобаның мақсаты – тандемдік қайталануларды, оның ішінде үшінші буынның толық геномдық секвенирлеудегі бастапқы деректерді идентификациялау мен вариабельділігін талдауға арналған ашық қол жетімді биоинформатикалық қосымшаны әзірлеу.

Күтілетін және қол жеткізілген нәтижелер:

Жоба аясында әр түрлі деңгейдегі дивергенциясы бар ұқсас тізбектерді идентификациялау үшін биоинформатикалық алгоритмдерді, сондай-ақ қосымшаның құрылымы мен интерфейсін әзірлеу кезінде бағдарламалау тілдері мен құралдары қолданады. Әзірленетін бағдарламалық жасақтамада алынған нәтижелердің дұрыстығын растау стандартты зертханалық молекулалық-генетикалық әдістермен, соның ішінде прокариоттар мен эукариоттардың геномдарының толық геномдық секвенирленуі және капиллярлық электрофорез арқылы тандемдік қайталануларды дифференциациялау әдістерімен жүзеге асырылады. Жүзеге асырылатын жобаның **негізгі нәтижесі** толық геномдық секвенирлеудегі тандемдік қайталануларды идентификациялауға арналған ашық қол жетімді және пайдаланушы интерфейсі бар бағдарламалық қамтама болады.

Бағдарламалық жасақтама тандемдік қайталанулармен мақсатты локустардың әртүрлілігін, соның ішінде толық геномдық секвенирлеудің бастапқы деректерінде идентификациялауға және идентификацияланған нұсқаларға статистикалық талдау жасауға мүмкіндік береді.

Жобаны іске асыру барысында Web of Science базасында импакт-фактор бойынша 1 (бірінші) немесе 2 (екінші) квартильдерге кіретін және (немесе) Scopus базасында CiteScore бойынша кемінде 65 (алпыс бес) процентилі бар рецензияланатын ғылыми басылымдарда кемінде 2 (екі) мақала және (немесе) шолулар немесе Web of Science базасындағы 1 (бірінші) квартильге немесе Scopus базасындағы CiteScore бойынша процентилі кемінде 95 (тоқсан бес) рецензияланатын ғылыми басылымда 1 (бір) мақала немесе шолу жарияланады. Барлық биоинформатикалық кодтар, скриптар тұрақты, ашық репозиторийлерде орналастырылады және еркін қол жетімді Github-та орналастырылады.

Зерттеу тобының мүшелері:

жоба жетекшісі – Исмаилова Айсулу Абжаппаровна, БҒҚ, PhD, қауымдастырылған профессор

ORCID: 0000-0002-8958-1846

Scopus/WoS (Hirsch Index = 3): Scopus Author ID: 56145830200

зерттеу тобы:

1) **Календарь Руслан Николаевич,** БҒҚ, б.ғ.к., биолог-генетик, Профессор (Биология), Генетика доценті (Хельсинки университеті)

ORCID: 0000-0003-3986-2460

Scopus/WoS (Hirsch Index = 34): ResearcherID: D-9751-2012

Scopus ID: 6602789279

2) **Бельдеубаева Жанар Толеубаевна,** ЖҒҚ, PhD

ORCID: 0000-0003-4056-6220

Scopus/WoS (Hirsch Index =2): Scopus Author ID: 56951278600

3) **Сатыбалдиева (Сатекбаева) Айжан Жанабековна**, ЖҒҚ, PhD

ORCID: 0000-0001-5740-7934

Scopus/WoS (Hirsch Index =2): Scopus Author ID: 56145597900

4) **Шевцов Владислав Александрович,** АҒҚ, техника ғылымдарының магистрі, «С. Сейфуллин атындағы Қазақ агротехникалық университеті» КеАҚ., «Ақпараттық жүйелер» кафедрасының «Үлкен деректерді талдау» ББ 2 курс докторанты

ORCID: 0000-0001-6202-2123

Scopus/WoS (Hirsch Index =3): Scopus Author ID: 57216896596

5) **Голенко Екатерина Сергеевна,** АҒҚ, техника ғылымдарының магистрі

ORCID: 0000-0002-4643-4571

6) **Вакансия**, ЖҒҚ, IT архитектор, бағдарламашы

7) **Вакансия**, ҒҚ, докторант

Әлеуетті пайдаланушыларға арналған ақпарат:

Әзірленген бағдарламалық қамтамасыз етуді **қолдану саласы:** биоинформатика, медициналық және ауылшаруашылық генетика, микроорганизмдердің генетикасы. **Бұл жобаның нәтижелері**, соның ішінде іргелі ғылымдар үшін үлкен маңызға ие. Бағдарламалық жасақтама тандемдік қайталануларды тиімді идентификациялауға және тандемдік қайталанулардың әртүрлілігі арасында адамның генетикалық ауруларымен және микроорганизмдердің генетикалық әртүрлілігімен және олардың патогенділігі арасында ассоциация орнатуға мүмкіндік береді. Жобаны іске асыру еліміздің жетекші ЖОО-да биоинформатика бағытын күшейтуге мүмкіндік береді және білімалушыларды мамандандыру мен кәсіптік бағдарлау үшін платформа әзірленеді.

Жоба бойынша 2023 жылғы алынған нәтижелер.

1) Белгілі табиғаты бар тандемдік қайталануы бар мақсатты локустары бар тізбектерді анықтауға, сондай-ақ жасырын қолтаңбасы бар және белгісіз табиғаты бар учаскелер үшін тандемдік қайталануларды болжауға мүмкіндік беретін кластар кітапханасы әзірленді. Анықталған тандемдер салыстырылатын генотиптердегі аллельдік нұсқаларды әрі қарай талдау және анықтау үшін олардың қолтаңбасына, қайталану сипатына және тандемдік блоктардың гетерогенділігіне қарай жіктеледі. Геномдық тізбектердегі қайталанулардың кез келген түрін анықтау үшін алгоритм және бағдарламалық код әзірленді. Сонымен қатар, қайталану талдауы NCBI генбанкінің толық геномдық тізбектерінде жүргізіледі. Қайталанатын тізбектер барлық геномдарда кездесетін функционалды барлық жерде кездесетін құрылымдық бірліктер болып табылады. Алайда, қайталанулардың әртүрлілігіне байланысты, олардың әрқайсысы ерекше қолтаңба мен құрылымға ие, оларды жіктеуді қиындатады. Бұл мәселені шешу үшін біз геномдық тізбектегі қайталанулардың кез келген түрін анықтауға арналған құралды әзірледік. Java эукариоттардың, саңырауқұлақтардың, микроорганизмдердің және алып вирустардың геномдарын қамтитын әртүрлі таксономиялық түрлердегі қайталануларды анықтау құралы және геномдық талдау нәтижелері мына жерде қол жетімді https://zenodo.org/records/8424601, және бастапқы код GitHub-та еркін қол жетімді https://github.com/rkalendar/Repeater.

2) Кейбір алгоритмдерді, соның ішінде жақын маңдағы сызықтық модельдерді, Кнут-Моррис-Пратт алгоритмін, Бойер-Мур алгоритмін, Рабин-Карп алгоритмін және суфикстік ағаштар алгоритмін қолдана отырып, тандемді қайталайтын реттілікті сәйкестендіру кодының тиімділігі тексерілді. Тестілеу кезінде тиімділікті бағалау ретінде келесі параметрлер таңдалды: әр алгоритмнің жұмыс жылдамдығы, табылған тандемді қайталау саны.

Тандемді қайталау тізбегін анықтау үшін кодтың тиімділігін тексеру процесінде келесі қадамдар орындалды:

- Сәйкестендіруді қажет ететін тандемді қайталаудың әртүрлі үлгілерін қамтитын биологиялық реттіліктің жасанды деректері дайындалды.

- Әр алгоритм дайындалған мәліметтерге қолданылатын бірқатар сынақтар өткізілді. Салыстыру кезінде суффикстік ағаштар қолданатын алгоритм тандемдік қайталануларды анықтау үшін алгоритмнің ең тиімді нұсқасы ретінде анықталды.

- Тандемді қайталау мен өнімділік үлгілерін табудың тиімділік критерийлері бойынша ең жақсы алгоритм анықталды.

- Суффикс ағаштарын қолданатын әдістерді қамтитын тандемді қайталауды іздеу үшін алгоритм әзірленді.

Нәтижесінде, суффикс ағаштары, жалпыланған суффикс ағаштары, суффикс ағаштарының көп жолды нұсқасы оңтайлы кеңістік пен уақытта есептеу биологияға қатысты есептерді шешу үшін пайдаланылуға болады.

Жобаны іске асырудың ағымдағы кезеңінде 3 ғылыми мақала дайындалды:

1) **Kalendar R**, Karlov GI 2023. Editorial: Mobile Elements and Plant Genome Evolution, Comparative Analyses and Computational Tools, Volume II. ***Frontiers in Plant Science***, 14: 1308536. DOI: 10.3389/fpls.2023.1308536.

https://www.frontiersin.org/articles/10.3389/fpls.2023.1308536/full

WoS IF_{2022}=6.627 Q1;

Scopus 88th percentile

https://www.scopus.com/sourceid/21100313905

2) Belyayev A, **Kalendar R**, Josefiová J, Paštová L, Habibi F, Mahelka V, Mandák B, Krak K 2023. Telomere sequence variability in genotypes from natural plant populations: unusual block-organized double-monomer terminal telomeric arrays. BMC Genomics 24, 572 (2023).

https://doi.org/10.1186/s12864-023-09657-y

WoS IF_{2022}=4.4 Q1;

Scopus 76th percentile

https://www.scopus.com/sourceid/21727

3) **Shevtsov V., Ismailova A., Beldeubayeva Zh., Satybaldiyeva A.,** Nurpeisova A. MLVA as a method of genotyping and algorithms for its implementation using genome-wide data. News of the National academy of sciences of the Republic of Kazakhstan. Physico-mathematical series. Volume 4. № 348 (2023). P. 300-312 https://doi.org/10.32014/2023.2518-1726.235