РЕЗЮМЕ НА НАУЧНИТЕ ТРУДОВЕ

на доц. д-р Веска Стефанова Ганчева

представени за участие в конкурс за заемане на академична длъжност ПРОФЕСОР
професионално направление: 5.3. Комуникационна и компютърна техника
научна специалност: Системи с изкуствен интелект
към катедра "Програмиране и компютърни технологии"
Факултет Компютърни системи и технологии, Технически университет – София
публикуван в ДВ бр. 28/02.04.2024

За участие в конкурса са представени научни трудове извън дисертационния труд и трудовете по конкурс за заемане на академичната длъжност „доцент" и регистрация на академична длъжност „доцент" в регистъра на НАЦИД: 50 научни публикации, вкл. 32 научни публикации в издания, които са реферирани и индексирани в Web of Science / Scopus; 2 учебника и 1 учебно пособие; 69 цитирания; 25 научноизследователски и образователни проекти; ръководство на 2 успешно защитили докторанта, разпределени по групи:

1. **Група В**
   1.1. Показател В.4. Хабилитационен труд – равностойни научни публикации (не по-малко от 10) в издания, които са реферирани и индексирани в световноизвестни бази данни с научна информация – 11.
2. **Група Г**
   2.1. Показател Г.7. Научни публикации в издания, които са реферирани и индексирани в световноизвестни бази данни с научна информация – 17.
   2.2. Показател Г.8. Научни публикации в нереферирани списания с научно рецензиране или в редактирани колективни трудове – 18.
3. **Група Д**
   3.1. Показател Д.12. Цитирания или рецензии в научни издания, реферирани и индексирани в световноизвестни бази данни с научна информация или в монографии и колективни томове – 58.
   3.2. Показател Д.14. Цитирания или рецензии в нереферирани списания с научно рецензиране – 11.
4. **Група Е**
   4.1. Показател Е.17. Ръководство на успешно защитил докторант (n е броят съръководители на съответния докторант) – 2.
   4.2. Показател Е.18. Участие в национални научни или образователни проекти – 14.
   4.3. Показател Е.19. Участие в международни научни или образователни проекти – 8.
   4.4. Показател Е.20. Ръководство на национален научен или образователен проект – 1.
   4.5. Показател Е.22. Привлечени средства по проекти, ръководени от кандидата – 1.
   4.6. Показател Е.23. Публикуван университетски учебник – 2.
   4.7. Показател Е.24. Публикувано университетско учебно пособие – 1.
   4.8. Показател Е.29. Ръководство на научен проект – 2.
5. **Група Ж**
   5.1. Показател Ж.30. Хорариум на водени лекции за последните три години – 722 часа.
6. **Група З**
   6.1. Показател З.31. Научни публикации в списания с импакт фактор (IF на Web of Science) и/или с импакт ранг (SJR на Scopus) – 4.

Хабилитационен труд - научни публикации (не по-малко от 10) в издания, които са реферирани и индексирани в световноизвестни бази данни с научна информация

Представените 11 научни публикации са свързани в обединяваща тема **“Интелигентни методи и средства за обработка на биомедицински данни**”. Трудовете са публикувани в реферирани и индексирани в световноизвестни бази данни с научна информация Scopus / Web of Science след получаване на образователната и научна степен „доктор” и заемане на академична длъжност „доцент“, в рецензирани научни издания: международни списания с IF/SJR (8) и сборници на международни конференции в чужбина (3).

В представените за рецензиране научни трудове са разгледани проблеми, свързани с използването на техники, алгоритми и средства от областта на изкуствения интелект за компютърни симулации в биоинформатиката и обработката на медицински изображения. Развитието на технологиите за генериране на биомедицински данни – секвенатори, които генерират генетични данни и средства за образна диагностика – компютърно томографските изображеня, изображенията от ядрено-магнитен резонанс, многослойните микроскопски изображения при клетъчния анализ и др. водят до натрупване на голям обем разнородни данни. През последното десетилетие сме свидетели на експлозия в количеството на наличните данни в областта на биоинформатиката и медицината. Количеството на данните става толкова голямо, че традиционните платформи и методи за анализ на данни вече не могат да отговорят на необходимостта от бързо изпълнение на задачите за анализ на данни и извличане на знания в науките за живота. Изкуственият интелект и машинното обучение, включително невронните мрежи, все повече навлизат в областта на биоинформатиката, здравеопазването и медицината. Предизвикателство при анализа на данни е да се предложи интегриран и съвременен достъп до прогресивно нарастващия обем данни, както и ефективни методи и алгоритми за тяхната обработка. Проведените изследвания са свързани с разработка на интегрирана платформа, състояща се от набор от софтуерни инструменти за автоматизиране на изчислителния процес при провеждане на научни експерименти и прилагане на интелигентни решения за управление и извличане на знания от биомедицински данни. Разработени са методи и алгоритми за анализ на биомедицински данни, базирани на математическо моделиране и синтез на метаданни, както и на машинно обучение, като са развивани и оптимизирани, за да осигурят високо качество на анализ, намалена изчислителна сложност, възможност за паралелна обработка. Резултатите са повишена ефективност при анализ на големи масиви биомедицински данни.

B.4.1. **Gancheva V.**, Stoev H. Optimization and Performance Analysis of CAT Method for DNA Sequence Similarity Searching and Alignment. *Genes*. 2024; 15(3):341, **IF=3.5 (2022) / SJR=0.817 (2023) / Q2, Scopus / WoS**

Публкацията представя нова версия на алгоритъма за подравняване на двойки ДНК последователности, базиран на предложения от авторите нов метод, наречен CAT, където се взема под внимание зависимостта с предишно съвпадение и най-близкия съсед, за да се увеличи уникалността на CAT профила и да се намалят възможните колизии, т.е. две или повече последователности с еднакви CAT профили. Това прави предложеният алгоритъм подходящ за по-бързо намиране на точното съвпадение на конкретна ДНК последователност в голям набор от ДНК данни. За да се даде възможност за използване на профилите като метаданни за последователностите, CAT профилите се генерират еднократно преди записването на данните в базата данни. Предложеният алгоритъм се състои от два основни етапа: изчисляване на CAT профил в зависимост от избраните бенчмарк последователности и сравнение на ДНК последователности

чрез използване на изчислените CAT профили. Подобренията в генерирането на CAT профили са подробно описани, актуализирани са блок-схемите, псевдокода, данните в таблиците и фигурите според предложената нова версия и получените експериментални резултати. Експериментите са проведени с помощта на новата версия на метода CAT за подравняване на ДНК последователности и различни набори от данни. Представени са нови експериментални резултати относно колизиите, скоростта и ефективността на предложената нова реализация. Експериментите, свързани със сравнението на производителността с Needleman–Wunsch са изпълнени с новата версия на алгоритъма, за да се потвърди, че има същата производителност. Извършен е анализ на производителността на предложения алгоритъм, базиран на метода CAT в сравнение с алгоритъма на Кнут–Морис–Прат, който има сложност O(n) и се използва широко за търсене на биологични данни. Изследвано е въздействието на предишните зависимости на съвпадението върху уникалността на генерираните CAT профили. Експерименталните резултати от подравняването на ДНК последователности показват, че предложеният алгоритъм, базиран на CAT метод, показва минимално отклонение, което може да се счита за незначително, ако такова отклонение се счита за допустимо в полза на подобрената производителност. Трябва да се отбележи, че производителността на CAT алгоритъма по отношение на времето за изпълнение остава стабилна, незасегната от дължината на анализираните последователности. Следователно, основното предимство на предложения подход се крие в неговите възможности за бърза обработка при мащабно подравняване на последователности – задача, за която традиционните точни алгоритми биха изисквали значително повече време за изпълнение. Подходът на предварително изчисление на метаданни и прилагане на принципа на трилатерация осигурява решение на проблема с бавното подравняване и търсене на сходство на биологични данни. Модифицирането на последователностите от бенчмаркове и начинън, по който профилите се изчисляват и сравняват, води до резултата от сравнението. Това прави подхода регулируем до желаното ниво на точност. Експериментите подчертават ефективността на предложения алгоритъм и неговия потенциал за значително ускоряване на процеса на подравняване на ДНК последователности чрез използване на прецизните CAT профили. Актуализираният алгоритъм обещава да бъде ценен инструмент в биоинформатиката, предлагащ по-бързи и по-надеждни средства за обработка на огромните и нарастващи хранилища на генетични данни.

B.4.2. **Gancheva V.,** Stoev H., An Algorithm for Pairwise DNA Sequences Alignment. Bioinformatics and Biomedical Engineering. *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol 13919, 2023, **Scopus, SJR = 0.606 (2023), Q3**

Предложен е нов алгоритъм за подравняване на ДНК последователности, базиран на предложения от авторите CAT метод, състоящ се от алгоритъм за изчисляване на CAT профил спрямо избраните бенчмарк последователности и алгоритъм за сравняване на две последователности, базиран на изчислените CAT профили. Определят се стъпките на изпълнение, входовете и изходите. Проектирана и разработена е софтуерна имплементация на предложения метод за подравняване на биологични последователности CAT. Проведени са експерименти с различни набори от данни за подравняване на ДНК последователности въз основа на метода CAT. Направен е анализ на експерименталните резултати по отношение на колизии, скорост и ефективност. Предложеният нов метод за подравняване на ДНК последователности е експериментално верифициран. Установени са три постоянни репера за прилагане на трилатерацията, което създава постоянна секвенция-фаворит, т.е. независимо от записите в базата данни остава същата когато наборът се промени. Тъй като установените бенчмарк последователности са постоянни (т.е. не зависят нито от данните, нито от техния брой), това позволява сравненията да се правят в самото начало – когато последователностите са записани в базата данни и това да бъде информация-метаданни, придружаваща всяка последователност. По този начин няма нужда да се сравняват последователности по време на търсене (най-бавната операция), като вместо това се сравняват само метаданните, генерирани при въвеждането на данните. Сравненията при търсене с метода CAT са

сведени до минимум и имат постоянна сложност на алгоритъма O(24), което помага за оптимизиране на търсенията в големи набори от биологични данни и го прави подходящ за внедряване като първа стъпка в по-прецизни алгоритми като FASTA. Въз основа на CAT профилите, последователностите могат да бъдат организирани в йерархична структура за съхранение, която да се използва като база данни за съхранение на биологични данни в системи, оптимизирани за търсене. Публикацията представя резултатите от разработената програмна реализация на предложения метод CAT за подравняване на биологични последователности. Проведени са експерименти с различни набори от данни за подравняване на ДНК последователности, като се използва базираният на триплет CAT метод. Направен е анализ на експерименталните резултати. Анализът на експерименталните резултати, получени чрез подравняване на последователности, показва малко отклонение на предложения алгоритъм, базиран на метода CAT, което може да бъде игнорирано, ако това отклонение е приемливо за сметка на производителността. Времето за изпълнение на алгоритъма на Needleman-Wunsch се увеличава с увеличаване на дължината на последователностите. Времевата ефективност на CAT алгоритъма остава постоянна, независимо от дължината на последователностите. Следователно предимството на предложения метод е бързата обработка при подравняването на големи последователности, за които изпълнението на точните алгоритми отнема много време.

B.4.3. **Gancheva V.**, Jongov T., Georgiev I. Medical X-Ray Image Classification Method Based on Convolutional Neural Networks, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol 13920, pp. 225-244, 2023, **Scopus, SJR = 0.606 (2023), Q3**

Изкуственият интелект и машинното обучение, включително конволюционните невронни мрежи, все повече навлизат в областта на здравеопазването и медицината. Целта на изследването е да се оптимизира процесът на обучение на конволюционни невронни мрежи чрез предварителна обработка на рентгенови изображения. Представен е модел за оптимизиране на цялостната архитектура на класифицираща конволюционна невронна мрежа от рентгенови снимки на гръдния кош чрез намаляване на общия брой конволюционни операции. Експерименталните резултати от изследването доказват успешното приложение на процеса на оптимизация върху обучението на класификационни конволюционни мрежи, като оптимизацията не влияе върху точността на обучените модели. Има значително намаляване на времето за обучение на всяка епоха в оптимизираните конволюционни мрежи. Оптимизацията е от порядъка на 25% за мрежата с размер на входния слой 124 × 124 и около 27% за мрежата с размер на входния слой 122 × 122. Методът може да се приложи във всяка област на класификация на изображения. В същото време няма значително отклонение в стойности на загубите и точността на данните за обучение на трите вида невронни мрежи. Стойностите на загубите и точността на валидиращите данни показват значителни вариации, които не дават значително предимство на нито една от невронните мрежи, а по-скоро са произволни. За провеждане на изследването са използвани три набора от данни – един набор от данни за рентгенови маски и два набора от данни за класификация на изображения. Изображенията са във формат jpeg или png със загуба на качество и не са с перфектно качество, но са подходящи за доказване на целта на изследването. Настоящото изследване се прилага за сегментиране и класифициране на рентгенови изображения на белия дроб, но методът може да се приложи във всяка област на класификация на изображения, в която информативните области на изображението са групирани и подлежат на сегментиране. От голямо значение е етапът на предварително сегментиране и анализ на активните области, при който се изследва разпределението на широчините и височините на получените активни области. Това разпределение е необходимо за определяне на ефективността на описания оптимизационен модел.

B.4.4. **Gancheva V.**, Georgiev I. A Scalable Healthcare Data Science Framework Based on Service-Oriented Architecture, In Proc. of International Conference on Research in Education and Science, May 18-21, 2023, Cappadocia, Turkiye, pp. 2525-2536, **Scopus, SJR=0.106 (2023)**

Целта на изследването е да предложи концептуален модел и архитектура на ориентирана към услуги мащабируема рамка, осигуряваща прилагането и валидирането на методи и алгоритми за интегриране, управление, анализ и визуализация на биомедицински данни и внедряване на научни изследвания за нуждите на прецизната медицина. Архитектурата на системата за анализ на големи биомедицински данни и откриване на полезни знания от данните се състои от следните компоненти: източници на биомедицински данни, съхранение на данни, интегриране и предварителна обработка на данни, поток от данни в реално време, обработка на потоци, съхранение на аналитични данни, моделиране и анализ на данни, и визуализация на резултатите. Използват се технологии за паралелна обработка и стрийминг. Множество данни могат да бъдат актуализирани в хранилището на данни за изследвания за анализ и визуализация. Проектирана е feed-forward изкуствена невронна мрежа, предназначена за анализ на данни, като по време на процеса на обучение входните данни се разделят на данни за обучение и данни за тестване. Определена е грешката на обучение и нейното разпределение върху теглата на невроните в мрежата. Като експериментални данни е използван намален набор от статистически записи, свързани с анализа на сърдечно-съдовите заболявания. Оригиналната база данни съдържа 76 атрибута, като 14 от тях са използвани за изследването. Данните са разделени в съотношение 0,8 към 0,2. 80% от данните са използвани за обучение на невронната мрежа, а останалите 20% за тестване на обучената мрежа. Изчислената точност се увеличава с увеличаване на епохите и е по-висока за данните за обучение и по-ниска за данните от теста за валидиране. Така трениранят модел може да бъде запазен и зареден на друга система, както и достъпен за преглед на стойностите на теглото. Обученият модел се прилага в системата за изчисляване на нови входни параметри, които не са били използвани нито при обучение, нито при валидиране.

B.4.5. **Gancheva V.** Platform for Big Biomedical Data Streams Management and Analytic, *International Journal of Circuits, Systems and Signal Processing*, pp. 580 - 588, ISSN 1998-4464, **Scopus, SJP = 0.156 (2020), Q4**

В публикацията е представена платформа за управление и анализ на биомедицински данни. Целта е да се предложи интелигентно решение като интегрирана, мащабируема среда за разработка на работни процеси, състояща се от набор от софтуерни инструменти за автоматизиране на изчислителния процес при провеждане на научни експерименти. Предложената платформа има за цел интелигентно управление на данни, анализ и визуализация. Предимствата в управлението на данни, анализа, откриването на знания и визуализацията дават възможност на учените да постигнат нови научни резултати. В резултат на това изследователската работа е насочена към разработване на компютърно подпомагана диагностична система за решаване на проблеми в областта на прецизната медицина. Представен е алгоритъм за прогнозиране на рак на гърдата, базиран на машинно обучение. Изследователските техники следват процеса на обработка за откриване на полезни знания от колекция от данни и обхващат следното: предварителна обработка на данни; откриване на знания и вземане на решения; включване на резултати и интерпретиране на точни решения от наблюдаваните резултати. Четири алгоритъма за машинно обучение за класифициране на рак на гърдата са избрани и оценени експериментално: Random Forest, kNN, Logistic Regression и SVM.

B.4.6. **Gancheva V.**, Borovska P. SOA Based System for Big Genomic Data Analytics and Knowledge Discovery, Proceedings of 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), DOI**:** 10.1109/IDAACS.2019.8924370, **Scopus/ Web of Science**

В публикацията е предложена система, базирана на архитектура ориентирана към услуги (SOA), за адаптивно откриване на знания и вземане на решения, базирани на анализ на големи геномни данни. Архитектурата на системата се състои от уеб услуги за интегриране на данни, предварителна обработка на големи потоци от данни, откриване на знания въз основа на анализ на геномни данни, интерпретация на знания и визуализация на резултатите. Проектираната системна архитектура се състои от уеб услуги за: (1) търсене и интегриране на разнородни данни от различни източници на данни и в различни формати; (2) подготовка, почистване, филтриране и избор на данни; (3) обработка и анализи на данни и (4) представяне на знания и визуализация на резултатите. Бъдещата работа е да се правят експерименти чрез системата в областта на молекулярната биология. Спектърът от казуси включва идентифициране на регулаторни елементи в секвенирани геноми и прогнозиране на вида и злокачествеността на рака на гърдата. Това ще позволи бърза обработка на данни от клинични наблюдения и лабораторни анализи и сравнение с натрупаните до момента налични данни в подкрепа на прецизната медицина.

B.4.7. **Gancheva V.**, Georgiev I. Software Architecture for Adaptive In Silico Knowledge Discovery and Decision Making Based on Big Genomic Data Analytics, *AIP Conference Proceedings 2172*, 090009 (2019), International Conference on Application of Mathematics in Engineering and Economics (AMEE'19), AIP Conference Proceedings, **Scopus/ Web of Science, SJR = 0.190 (2019)**

В тази публикация е представена софтуерна архитектура за адаптивно откриване на знания, базирана на анализ на големи геномни данни. Софтуерната архитектура се състои от слоеве за интегриране и предварителна обработка на данни, база данни/сървър за съхранение на данни, машина за откриване на данни, оценка на шаблони и графичен потребителски интерфейс. Архитектурата на големите геномни данни се състои от източници на данни, съхранение, интеграция и предварителна обработка, реален поток от данни, обработка на потоци, съхранение на аналитични данни, анализ и докладване. Представен е алгоритъм за прогнозиране на рак на гърдата, базиран на машинно обучение за обработка и анализ на големи геномни данни и откриване на знания по отношение на персонализирано лечение. Предложеният алгоритъм за класификация на рака на гърдата се прилага чрез използване на метода на Stochastic Dual Coordinate Ascent (SDCA) и база данни за рак на гърдата в Уисконсин. Представени и обсъдени са експерименталните резултати. Целта на изследването е да се приложи софтуерната архитектура за анализа на големи геномни данни чрез практически експерименти за конкретен казус, идентифициращ регулаторни генетични елементи в секвенирани геноми и прогнозиране на типа и злокачественост на заболяването рака на гърдата. Това позволява бърза обработка на данни от клинични наблюдения и сравнение с наличните данни, натрупани досега в подкрепа на прецизната медицина. Предложената софтуерна архитектура се основава на машинно обучение и процедури за генериране на модели и правила, съобразени с целта на научното изследване. Предложен е интегриран подход за подпомагане на откриването на знания, базиран на адаптивно машинно обучение и адаптивни процедури за генериране на правила според целта на научното изследване. Предимството на предложената рамка е автоматичното генериране на хипотеза и опции за вземане на решения въз основа на анализ на набор от данни за обучение, като проверката и валидирането се извършват чрез набор от данни за тестване и експертиза на изследователи в съответната област.

B.4.8. **Gancheva V.**, Stoev H. DNA Sequence Alignment Method Based on Trilateration, Bioinformatics and Biomedical Engineering, *Lecture Notes in Computer Science*, 2019, vol. 11466, Springer, Cham, pp. 271-283, **Scopus/ Web of Science, SJR = 0.427 (2019), Q2**

Ефективното сравнение на последователности от биологични данни е важна и предизвикателна задача в биоинформатиката. Самият процес на подравняване на последователности е начин за подравняване на ДНК последователности, за да се идентифицират подобни области, които могат да имат следствие от функционални, структурни или еволюционни връзки между тях. Целта на тази публикация е да представи нов ефективен и унифициран метод за подравняване на ДНК последователности на базата на метода на трилатерация. Този метод предлага решения на три основни проблема при подравняването на последователности: създаване на постоянна реперна последователност, намаляване на броя на сравненията и обединяване/ стандартизиране на реперната последователност чрез дефиниране на последователности за сравнение. Дефинирани са три постоянни бенчмарка за прилагане на трилатерация, които създават постоянна реперна последователност, т.е. не зависи от данните в базата данни и остава същата при промяна. Това позволява да се правят сравнения в самото начало – по време на въвеждане на последователностите в базата данни и може да се запише като метаданни към всяка последователност. По този начин няма нужда да се прави сравнение на последователностите по време на търсенето, а вместо това ще се сравняват само метаданните. Чрез установяване на еталонни последователности е решен проблемът с унифицирането/ стандартизирането на реперната последователност за всички последователности, използвайки описания алгоритъм за сравнение. Изчисленията в предложения алгоритъм, са относително прости и бързи, което го прави подходящ за използване като първа стъпка в алгоритми за подравняване на биологични последователности.

B.4.9. Borovska P., **Gancheva V.**, Georgiev I. Platform for Adaptive Knowledge Discovery and Decision Making Based on Big Genomics Data Analytics, Bioinformatics and Biomedical Engineering, *Lecture Notes in Computer Science*, 2019, vol. 11466. Springer, Cham, pp. 297-308, **Scopus/ Web of Science, SJR = 0.427 (2019), Q2**

През последните години изследователи и анализатори по света определят големите данни като революция в научните изследвания и една от най-обещаващите тенденции, която даде тласък на интензивното развитие на методи и технологии за тяхното изследване и доведе до появата на нова парадигма за научни изследвания Data-Intensive Scientific Discovery (DISD). Статията представя платформа за адаптивно откриване на знания и вземане на решения, базирани на анализ на големи данни и съобразени с целта на научното изследване. Основното предимство е автоматичното генериране на хипотези и варианти за решения, тъй като проверката и валидирането се извършват с помощта на стандартни набори от данни и експертиза на учени. Платформата е реализирана на базата на мащабируема рамка и научен портал за достъп до базата знания и софтуерните инструменти, както и възможности за споделяне на знания и трансфер на технологии. Уеб порталът предоставя услуги за достъп и извличане на знания от биологични данни и изпълнение на паралелни софтуерни приложения за анализ на големи геномни данни. Представен е интегриран подход за подпомагане на откриването на знания и вземането на решения, базиран на анализ на големи данни, адаптивно машинно обучение и адаптивни процедури за генериране на правила според целта на научното изследване. Бъдещата работа е да се направят in silico експерименти на платформата, базирана на големи анализи на геномни данни за научни изследвания в областта на молекулярната биология. Спектърът от казуси, които се разследват, включва идентифициране на регулаторни елементи в секвенирани геноми и прогнозиране на вида и злокачествеността на рака на гърдата. Това ще позволи бърза обработка на данни от клинични наблюдения и лабораторни анализи и сравнение с натрупаните досега налични данни в подкрепа на прецизната медицина.

B.4.10. Borovska P., **Gancheva V.**, Georgiev I., Ivanova D., Hybrid Parallel Multiple Sequence Alignment Based on Artificial Bee Colony on the Supercomputer JUQUEEN, Proceedings of International Conference on Electrical Engineering and Computer Science (EECS), Bern, Switzerland, 2017, pp. 47-51, ISBN: 978-1-5386-2086-1, DOI**:** 10.1109/EECS.2017.18, **Scopus/ Web of Science**

Фокусът на публикацията е върху изследване на производителността и подобряване на софтуера за подравняване на множество биологични последователности MSA_BG на суперкомпютъра BlueGene/Q JUQUEEN. За тази цел са проведени научни експерименти в областта на биоинформатиката, като са използвани като казус последователности на грипния вирус. Паралелният софтуер MSA_BG за подравняване на множество последователности е пренесен и настроен на суперкомпютъра Blue Gene/Q JUQUEEN. Целите на изследването са оптимизиране на кода, мащабиране, профилиране и оценка на производителността на софтуера MSA_BG. За тази цел е разработена хибридна паралелизация чрез MPI/OpenMP върху кода за MPI и са демонстрирани предимствата на този подход чрез резултатите от сравнителни тестове, извършени на JUQUEEN. Паралелната производителност е изследвана и оптимизирана чрез сравнителни тестове и профилиране. Експерименталните резултати показват, че хибридната паралелна реализация осигурява значително по-добра производителност от оригиналния код. Внедряването на хибридно паралелизиране чрез MPI/OpenMP на MSA_BG намалява общото време на изпълнение на MPI версията на приложението, тъй като позволява да се заеме напълно капацитета на процесора на JUQUEEN възел с нишки. Трябва също да се отбележи, че изпълненията, използващи хибридното внедряване, доведоха до по-добро качество на подравняването на последователностите поради допълнителната произволност, произведена от генератора на Mersenne Twister. Оценката на производителността и анализите показват, че внедряването на хибридна паралелна програмна имплементация на алгоритъма MSA_BG се мащабира добре, тъй като броят на ядрата се увеличава и е добре балансиран както по отношение на работното натоварване, така и по отношение на размера на машината. Оптимизираният код е универсален и може да се прилага и за други подобни изследователски проекти и експерименти в областта на биоинформатиката. Софтуерът MSA_BG позволява на изследователите да провеждат своите експерименти и симулации с много големи масиви данни.

B.4.11. Borovska P., **Gancheva V.**, Landzhev N. Massively Parallel Algorithm for Multiple Biological Sequences Alignment, Proceedings of the IEEE International Conference on Telecommunications and Signal Processing (TSP), Rome, Italy, ISBN 978-1-4799-0402-0, pp. 638-642, DOI**:** 10.1109/TSP.2013.6614014, **Scopus/ Web of Science**

Обработката на биологичните последователности е ключ за молекулярната биология. Тази научна област изисква мощни изчислителни ресурси за изследване на големи набори от биологични данни. Подравняването на множество последователности е широко използван метод за обработка на биологични последователности. Целта на този метод е подравняване на ДНК и протеинови последователности. Статията представя иновативен паралелен алгоритъм MSA_BG за множествено подравняване на биологични последователности, който е силно мащабируем. Проектираният алгоритъм MSA_BG е итеративен и се основава на концепцията за метаевристика на изкуствената пчелна колония (ABC) и концепцията за корелация на алгоритмични и архитектурни пространства. Конструирана е метафората на метаевристиката ABC и са дефинирани функционалностите на агентите. Проектиран е концептуален паралелен модел на изчисление. Конструирана е алгоритмичната рамка на проектирания паралелен алгоритъм. Изборът на алгоритъма ABC се основава на факта, че по същество това е хибридна метаевристика - комбинация от методи, базирани на популации (скаутите генерират едновременно няколко възможни решения) и метод, базиран на траектории (наетите пчели извършват локални търсения около решенията на скаутите, стремейки се да подобрят качеството на решенията). Алгоритъмът MSA_BG има йерархична структура, която позволява спазване на принципа на локалност (независими изчисления) и много

висока скалируемост (кошери и рояци), така че се очакват високоефективни реализации за петафлопс суперкомпютри. Представени са оценка на паралелната производителност и профилиране на подравняване на множество последователности на базата на алгоритъм MSA_BG на суперкомпютър BlueGene/P. Казусът изследва вирусните нуклеотидни последователности и открива консенсусни мотиви и променливи домейни в различните сегменти на грипния вирус. Паралелните параметри на производителността като време за изпълнение, време за ускоряване и профилиране са оценени експериментално. Анализите на оценката на производителността и профилирането показаха, че паралелната система е добре балансирана както по отношение на работното натоварване, така и по отношение на размера на машината, с изключение на процеса с ранг 0, който е най-силно използван поради разпространение на данни към всички други процесори, комуникация и синхронизация.

Публикации извън хабилитационния труд

Г.7.1. Trendafilov I., Gancheva V. Neuromorphic Assisted Sensor Grids, XXXII International Scientific Conference Electronics, 2023, DOI: 10.1109/ET59121.2023.10279437, Scopus

Конкретен проблем при използването на техники за изкуствен интелект в сензорната мрежа е високата консумация на енергия. Дистанционните сензори обикновено са ограничени от количеството налична мощност, поради което общата цел е да се сведе до минимум, като същевременно се запазят точните резултати от експериментите. Използвайки нововъзникващи технологии като импулсни невронни мрежи и невроморфен хардуер, може да се създадат сензорни мрежи, които извършват обработка на данни със запазване на ниска консумация на енергия. В публкацията е демонстриран отдалечен възел с интегриран визуален сензор, който работи с по-малко от 10 mW енергия, като същевременно извършва непрекъснат мониторинг на сцената, откриване и класифициране на обекти. Системата има интелигентно управление на захранването и възможност за безжично изпращане на данни.

Г.7.2. Trendafilov I., Gancheva V. Neuromorphic Neurons and Networks for Artificial Intelligence Built Using Temporal Space Calculations, XXXII International Scientific Conference Electronics, 2023, DOI: 10.1109/ET59121.2023.10279371, Scopus

Създадена е импулсна невронна мрежа, която изчислява теглата на мрежата във времевото измерение. Такава мрежа може да се използва за изкуствен интелект и дълбоко обучение. Предложени и демонстрирани са схеми, изпълняващи блокове за изграждане на такава мрежа и след това модел за обучение. Това позволява създаването на ефективни детектори на Hessenstein-Reichardt, наблюдавани в мрежи за откриване на движение в природата. Предложената мрежа позволява алгоритмично преконфигуриране през мрежовите слоеве. Проектираната нова импулсна невронна мрежа опростява системните параметри и намалява размерността на сигнала до двоично сигнализиране. Този тип мрежи трябва да бъдат стабилни и лесни за внедряване в невроморфен хардуер, позволяващ изграждането на много големи мрежи с милиарди неврони. Алгоритъмът за обучение позволява преконфигуриране на мрежата, поведение, наблюдавано в природата. Необходими са допълнителни изследвания, но работната хипотеза е, че такава мрежа може да се използва за изпълнение на всяка задача, която изисква предишно състояние на всеки неврон в мрежата. В публикацията е демонстрирана техническата жизнеспособност на метода, следователно може да продължат изследванията със софтуерни симулации на такива мрежи. Демонстрирано е

използване на заряда като междинна количествена променлива. Използвани са кондензатори за съхраняване на електрическия заряд, но в бъдеща работа се планира проучване и използване на мемристори в тази роля.

Г.7.3. Trendafilov I., **Gancheva V.** Hassenstein-Reichardt Detector Using Controllable Single Pulse Time-Delay Circuit for Neuromorphic Hardware, International Scientific Conference Computer Science, 2023, DOI: 10.1109/COMSCI59259.2023.10315865, **Scopus**

Вдъхновени от визуалната система на плодовата мушица, е създаден общ градивен елемент за невроморфен хардуер, който е жизненоважен за трето поколение невронни мрежи, като позволява закъснението да бъде параметризирано по ефективен начин. Предимството на предложеното решение е възможността за изграждане и тестване на нови мрежи с помощта на динамичен подход в изкуствения интелект. Проектирана е схема, която създава линия на забавяне на импулса с контролируеми времеви параметри, която може да се използва за изграждане на детектори на Hassenstein-Reichardt и интегриране в невроморфен хардуер, работещ с импулсни невронни мрежи. Параметрите могат да се променят по време на обучение на невронната мрежа. Схемата е реализирана чрез използване на два кондензатора, всеки сдвоен с контролируем източник на напрежение. Това осигурява два независими времеви параметъра. Първият кондензатор има заряд, пропорционален на дължината на входния импулс, докато вторият задава закъснението между входа и изхода.

Г.7.4. **Gancheva V.** Application of Machine Learning Techniques for Software Anomaly Detection, International Conference on Applied Mathematics & Computer Science (ICAMCS), Lefkada Island, Greece, August 8-10, 2023, IEEE Catalog Number: CFP23T98-ART, ISBN: 979-8-3503-2426-6, DOI: 10.1109/ICAMCS59110.2023.00016, **Scopus**

През последните години все по-голямо разнообразие от платформи и софтуерни програми използват набори от данни, съхранявани в хранилища с отдалечен достъп. В резултат на това, наборите от данни са по-уязвими за злонамерени атаки и съответно мрежовата сигурност придобива все по-голямо значение като изследователска тема. Използването на системи за откриване на проникване е добре позната стратегия за защита на компютърните мрежи. Изследването, представено в тази публкация, предлага хибриден метод за откриване на аномалии, който съчетава методи, базирани на правила и базирани на машинно обучение. Предимството на предложената методика е комбинацията от различни методи и алгоритми. Прилага се генетичен алгоритъм за изграждане на съответните правила. Анализът на основните компоненти се използва за извличане на съответните характеристики, насочени към подобряване на производителността. Предложеният метод за откриване на софтуерни аномалии се проверява експериментално чрез прилагане на три класификационни алгоритъма. Направен е анализ и оценка на получените резултати от гледна точка на точност и прецизност. Предложеното решение се използва за идентифициране и изследване на четири различни вида атаки в набор от данни за бенчмарк: Neptune, Ipsweep, Pod и Teardrop. Наборът от данни KDD Cup 1999, който отговаря на условието за използване на подходящи данни, се използва за емпирично валидиране на предложения метод. Наборът от данни за KDD Cup 1999 се състои от 41 функции, които са разбити на задължителни, трафик и аспекти на съдържанието, както и данни за обучение и тестове. Наборът от данни за KDD Cup 1999 съдържа приблизително пет милиона необработени точки от данни, като данните за атака представляват около 80% от тях. След тестване на характеристиките, посочени във фазата на

обучение, данните се класифицират в категории атаки и нормално поведение по време на фазата на машинно обучение. Четири "атакуващи" групи и една "нормална" категория включват тези статистики. Експериментите се извършват на базата на алгоритмите Support Vector Machine, Decision Tree и Naive Bayes и са насочени към точност и вероятност при анализа на набори от данни. Направеният анализ показва най-добри резултати в случая на класификационния алгоритъм на Naive Bayes и може да се приеме, че е най-надежден в сравнение с резултатите в случаите на Support Vector Machine и Decision Tree.

Г.7.5. Draganov I., Gancheva V. Optimizing the Non-local Means Filtering of CT Images. Medical Imaging and Computer-Aided Diagnosis. MICAD 2022. *Lecture* Notes *in Electrical Engineering*, vol 810. Springer, Singapore, https://doi.org/10.1007/978-981-16-6775-6_1, Scopus, SJR=0.147 (2022), Q4

В публикацията е предложена обща оптимизационна схема за контролните параметри на филтъра за нелокални средства (NLM). Това включва намиране на оптималната степен на изглаждане, размера на прозореца за търсене и размера на прозореца за сравнение за серия изображения от компютърна томография (CT). Всички те съдържат адитивен бял шум на Гаус (AWGN) с определена дисперсия и нулева средна стойност, като и двете са предварително неизвестни. Прилагането на процедурата за оптимизация върху един срез от CT пакета изглежда достатъчно ефективно за намиране на оптималните параметри на филтъра за останалите CT изображения. Положителни резултати се получават от филтриране на пълен набор от CT изображения от тялото на пациент и качеството на филтрирането е по-високо от това на филтрите на Gaussian и Average филтър. Експерименталните резултати показват, че степента на изглаждане влияе върху качеството на реконструираните изображения. Увеличаването на този параметър води до насищане както на Peak Signal to Noise Ratio (PSNR), така и на Structural Similarity Index Measure (SSIM). Има минимална стойност за Degree of Smoothing (DoS), която може да се намери като оптимална в началото на зоната на насищане. Промяната на DoS няма значителен ефект върху времето за филтриране. Размерите на прозорците за търсене и сравнение също имат нелинеен ефект върху качеството на реконструираните изображения. И за двете има области на насищане във функциите PSNR и SSIM. Възможно е да се избере минималните размери на прозорците, така че да лежат в началото на зоната на насищане. По този начин те гарантират най-добро качество на изображенията при най-ниско изчислително време. Самото време за изчисление се увеличава монотонно с увеличаване на повърхността на прозореца за търсене и сравнение. Филтърът NLM осигурява по-добро качество на филтрираните CT изображения от филтрите на Gaussian и Average за широк диапазон от ниво на шум на AWGN. Времето за филтриране и на трите филтъра не зависи от нивото на шума. NLM филтърът е с повече от 2 порядъка по-бавен от другите два филтъра. Няма зърнеста структура в изображенията, филтрирани от NLM, но има малка загуба на контраст. Като бъдеща работа може да се предприеме оптимизиране на NLM филтъра като време за обработка.

Г.7.6. Ganheva V., Todorova V. Workflow for Medical Data Classification and Analysis, 6th International Symposium on Multidisciplinary Studies and Innovative Technologies, October 20-22, 2022, Ankara, Turkey, DOI: 10.1109/ISMSIT56059.2022.9932780, Scopus

Публикацията представя подход за автоматизирано извличане на знания и вземане на решения от медицински изображения чрез работен процес за предварителна обработка на входящи рентгенови изображения, анализ, класификация и оценка на резултатите. Проектираният алгоритъм за анализ на медицински рентгенови изображения се основава на машинно обучение и се състои от три основни фази: предварителна обработка на набори от данни за обучение и валидиране,

класификация на медицински изображения с помощта на алгоритми за машинно обучение като логистична регресия, Naive Bayes, SVM, оценка на модела. Разработен е работен процес за автоматизирана обработка и анализ на набори от данни на белодробни рентгенови изображения, съдържащи четири класа, и определяне на точността на класификация чрез изследване на параметрите за оценка на ефективността. Изчисляването на разстоянието се извършва чрез косинусово разстояние. Извършеният анализ показва предимството на резултатите от логистичната регресия, които се приемат за по-добри в сравнение с резултатите, получени от Naive Bayes и SVM. Атрибутът Category е избран като цел за класификацията. 66% от данните са избрани като набор от данни за обучение. Останалите данни се използват като набор от тестови данни.

Г.7.7. **Ganheva V.**, Vetova S. Approach and Concept of Workflow for Animal Husbandry Data Integration and Analysis, 30th National Conference with International Participation (TELECOM), 27 - 28 October 2022, Sofia, Bulgaria, pp. 1-4, DOI: 10.1109/TELECOM56127.2022.10017334, **Scopus/ Web of Science**

Представена е концепция за интегриране на данни и анализ на риска в животновъдното производство. Предложената структура на модела за интегриране и анализ на данни за животновъдството се състои от три слоя, всеки от които комбинира задачите, които трябва да бъдат изпълнени. Описана е архитектурата за интегриране на данни и нейните компоненти. Организацията на работния процес е представена детайлно, включително модулите и техните връзки, използвани за обмен на данни. На базата на представения работен процес са реализирани експерименти с използване на статистически данни за популацията на домашни животни. Анализът на получените резултати показва тенденцията в риска на животновъдната продукция. Представено е прилагането на концепцията на модела на работния процес за интегриране и анализ на данни в животновъдството и по-специално оценката на риска от изчезване на породи домашни животни. Моделът демонстрира способността за зареждане, обработка и анализ на 141 записа, първоначално дефинирани в оригиналния статистически CSV файл. Статистиката предоставя информация за анализ на риска от изчезване на избраните видове домашни животни по породи. Създаденият модел за обработка и анализ на статистически данни за оценка на риска от изчезване на породи животни позволява проследяване на тенденцията на растеж на дадена порода за определен период от време въз основа на предварително събрани статистически данни. Въз основа на автоматизираната обработка на данните и оценката на риска от изчезване и заключенията, като следваща стъпка могат да се предприемат мерки за защита на породите животни. Статистиката предоставя информация за анализ на риска от изчезване на избраните видове домашни животни по породи.

Г.7.8. Draganov I., **GanchevaV.** Unsharp Masking with Local Adaptive Contrast Enhancement of Medical Images, *Lecture Notes in Electrical* Engineering, vol 784. Springer, **Scopus, SJR = 0.147 (2022), Q4**

Предложен е обобщен алгоритъм за нерезкостно маскиране на медицински изображения, който приема като един от входните параметри висококонтрастно изображение, подложено на локално адаптивно подобрение на контраста. Изборът на оптимални стойности на броя на хистограмните контейнери, размерът на прозореца и интензитета на отсичане на долните и горните граници по итеративен начин е част от прилагането на Contrast Limited Adaptive Histogram Equalization (CLAHE). Представени са процедури за оптимизиране на алгоритмите за изравняване на хистограмата, регулиране на интензитета и контрастните ограничения, за да се намерят оптимални параметри за тях. Средният квадратичен контраст, резкостта и структурното сходство между изображение с увеличен контраст и оригиналното изображение играят ролята на целеви параметри. Експерименталните резултати разкриват по-високо качество на изходните изображения както по

отношение на средния контраст на съдържанието, така и по отношение на остротата. Постигнатото качество, както визуално, така и количествено, се сравнява с това от алгоритъма за адаптивно изравняване на хистограмите (AHE), ограничено разтягане на хистограмата и обикновено изравняване на хистограми, което доказва неговата приложимост. Тестовете с CT и рентгенови изображения потвърждават правдоподобността на предприетия подход и приложимостта на получените изображения за алгоритъма за маскиране на неравномерност, за да ги използва като вход. Ограниченото по контраст адаптивно изравняване на хистограмата дава по-детайлни и подобрени с контраст крайни изображения, последвани от изравняването на хистограмата и алгоритмите за настройка на изображението на цената на повече изчислително време. Нерезкостното маскиране в тази обща и лесна за изпълнение форма се смята за полезен инструмент за медицински цели. Алгоритъмът се счита за подходящ за обработка на редица видове изображения, като CT, рентгенови снимки и др.

Г.7.9. Draganov I., **Gancheva V.** Optimal Bilateral Filtering of CT Images, International Conference on Computational Science and Computational Intelligence (CSCI), 2021, pp. 1668-1672, DOI: 10.1109/CSCI54926.2021.00053, **Scopus/ Web of Science**

В публикацията е предложен оптимизационен алгоритъм за настройка на параметрите на билатерален филтър за обработка на изображения от компютърна томография, съдържащи адитивен бял Гаусов шум. Отношението пиков сигнал/шум Peak Signal to Noise Ratio (PSNR) и мярката за индекс на структурно сходство Structural Similarity Index Measure (SSIM) са целевите параметри, използвани по време на оптимизацията с поставена цел за намиране на техните максимуми. Работата на оптималната конфигурация на билатералния филтър се сравнява с резултатите от филтрирането на едни и същи изображения с Гаусов и осредняващ филтър. Получени са положителни резултати и предложената оптимизация се счита за приложима не само за изображения с компютърна томография, но и за изображения от ядрено-магнитен резонанс, мултиспектрални и хиперспектрални изображения и др. Обективните параметри за качество зависят както от покритието, така и от използвания обхват на интензитета, чиято зависимост става по-силна с увеличаването на дисперсията на наличния шум. Прилагането на процедурата за оптимизация върху отрязък от CT изображение и последващо филтриране на всички срезове осигурява ефективен начин за получаване на най-високо качество за целия набор. Бъдеща работа по тематиката ще разкрие неговата приложимост не само за различни видове изображения, но и по-нататъшно усъвършенстване на оптимизацията за различни видове шумове, като се използват подходящи адаптирани форми на филтъра.

Г.7.10. **Gancheva V.** Parallel Multithreaded Medical Images Filtering, International Conference on Computational Science and Computational Intelligence (CSCI), 2021, pp. 1788-1793, DOI: 10.1109/CSCI54926.2021.00338, **Scopus/ Web of Science**

Качеството на медицинските изображения е от първостепенно значение, гарантира качество на медицинската диагностика, лечение и качество на живот на пациента чрез средствата на здравеопазването или чрез използване на автоматизирани интелигентни системи за медицинска диагностика, лечение и наблюдение. Публикацията представя изчислителните предизвикателства при обработката на медицински изображения. Големите предизвикателства са да се предложат паралелни изчислителни модели и паралелни програмни реализации, базирани на алгоритмите за филтриране на медицински изображения. Предложен е работен процес за филтриране на медицински изображения, който включва следните филтри: контрол на яркостта, Laplace, Blur,

хоризонтален и вертикален филтър Sobel. Проектиран е паралелен изчислителен модел, базиран на двумерни филтри. Предложеният паралелен модел е верифициран чрез изпълнение на многонишкова паралелна програмна имплементация. Изследвана е ефективността на филтрите за медицински изображения, базирана на паралелната многонишкова програмна имплементация, прилагаща двуизмерни филтри върху даден списък от компресирани jpeg медицински изображения и генериране на изходни jpeg изображения за всеки тип приложен филтър. Проведени са редица експерименти за случая на набор от данни, състоящ се от 162 цялостни изображения на слайдове на проби от рак на гърдата (BCa), сканирани при 40x и различен брой нишки. Времето за изпълнение и ускоряването на параметрите на паралелната производителност са оценени експериментално. Оценката на производителността и анализите на мащабируемостта показват, че предложеният модел има добра мащабируемост.

Г.7.11. Ko S.-H., Gancheva V. An Approach for Parallel Reading in Multiple Sequence Alignment, International Conference Automatics and Informatics, ICAI 2020, DOI: 10.1109/ICAI50593.2020.9311347, Scopus

Предложен е подход за по-бързо четене на входни файлове за множествено подравняване на последователности чрез използване на MPI-I/O върху подмножество от MPI ядра. Идеята е да се позволи на подмножество от MPI ядра да извършва I/O операция и локално да се излъчва към отделни съседи, така че кодът да е по-малко чувствителен към стабилността на паралелната файлова система. Това се постига чрез създаване на редица подгрупи под глобален MPI комуникатор. Размерът на всяка подгрупа и размерът на буфера на всяка операция за четене се настройват чрез синтетичен бенчмарк. Ефективността на предложения подход е верифицирана, като е сравнен с традиционния начин на „последователно четене на файлове и глобално излъчване“ и е приложен към MPI версията на софтуера за множествено подравняване на последователности ClustalW. Разделен е общият брой MPI ранг на подгрупи и е оставен локалния главен на всяка група да извършва I/O операции, вместо да се отваря I/O протокол за всички рангове. Заедно с този размер на групата, размерът на данните за една инструкция за четене също влияе много върху I/O производителността. Проведени са сравнителни експерименти, за да се определят оптималните стойности на тези два параметъра. Изпълненията на бенчмарка показват, че най-добрата производителност се постига, когато размерът на групата е 1/4 или 1/8 от общия брой процесори и размерът на частта за четене е зададен като размер на файла. В тази конфигурация паралелният I/O превъзхожда серийния I/O 2 – 4 пъти при набори от данни от десетки или стотици мегабайта. При производствения цикъл на 8192 BlueGene/Q ядра, настоящият подход осигурява 6.8 пъти ускорение от оригиналното изпълнение на ClustalW-MPI. Предложеният дизайн на паралелен I/O интерфейс може да осигури голяма печалба в I/O производителността на много биоинформатични софтуери.

Г.7.12. Aleksieva-Petrova A., Gancheva V., Petrov M. Software Architecture for Adaptation and Recommendation of Course Content and Activities Based on Learning Analytics, Proceedings of International Conference on Mathematics and Computers in Science and Engineering, DOI: 10.1109/MACISE49704.2020.00010, Scopus/ Web of Science

Публикацията представя софтуерна архитектура за адаптиране и препоръчване на съдържанието и дейностите на учебен курс, базирани на анализ на обучението. Състои се от слой за приемане, слой за агрегиране, слой за съхранение и слой за обработка и анализ на големи данни. Представен е алгоритъм за прогнозиране на обучението на студентите, базиран на машинно обучение за обработка и анализ на данни и откриване на знания по отношение на основните дейности на обучаемия и учителя. Предложеният алгоритъм за класифициране на обучението на студентите е

реализиран с помощта на метода на осреднения перцептрон. Проведени са експерименти е са обсъдени поручените резултати. Целта на изследването е да се приложи софтуерната архитектура за анализ на обучението чрез практически експерименти за специфични казуси, идентифициращи елементи на събития в последователни дневници на дейностите на обучаемите и курсовете и прогнозиране на обучението на студентите, както и адаптиране и препоръчване на съдържанието на курса и дейностите, базирани на анализ на обучението. Обучението в реално време и анализът на големи данни, генерирани от модерни платформи за електронно обучение и образователни игри, за ориентирано към учащия адаптиране на технологично подобреното обучение е едно от основните предизвикателства. Системата помага за структуриране и съхранение на големи данни от хетерогенни източници както като LMS, така и като образователна игра; идентифициране на модели, като анализира поведението на обучаемите и позволява анализи на данни с описателни, предсказуеми и предписващи резултати. Експерименталният набор от данни е получен от системата за управление на обучението и съдържа 63774 екземпляра, характеризиращи се със 7 атрибута.

Г.7.13. Gancheva V. Knowledge Discovery Based on Data Analytics and Visualization Supporting Precision Medicine, International Conference on Mathematics and Computers in Science and Engineering, pp. 102 - 105, DOI: 10.1109/MACISE49704.2020.00024, Scopus/ Web of Science

Една цялостна система за прецизната медицина, която обхваща всички фази на откриване на данни, интегриране на данни, предварителна обработка на данни, изграждане на модели, съхранение на данни, анализ на данни и визуализация, може да бъде много полезна за учените в подкрепа на прецизната медицина. Софтуерната система има за цел интелигентно управление, анализ и визуализация на големи геномни данни и позволява на учените лесен, бърз и гъвкав подход за обработка на данните. Те могат да избират услугите, които желаят да бъдат изпълнени, да използват наличните набори от данни в базите данни или да въвеждат свои собствени данни, които да бъдат обработени. Разработено е софтуерно приложение за визуализация на биологични данни с цел тестване и валидиране на системата. Предлаганото приложение предоставя възможност за триизмерна визуализация на структурата на протеините или ДНК последователност, реализирана чрез OpenGL. Триизмерното моделиране на съответните макромолекули позволява да се получи ясна представа за сложността на обектите на атомно ниво. Сложните молекули могат да бъдат показани чрез използване на съвременни технологии за 3D моделиране.

Г.7.14. Gancheva V., Georgiev I. Multithreaded Parallel Sequence Alignment Based on Needleman-Wunsch Algorithm, Proceedings of 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), DOI: 10.1109/BIBE.2019.00037, Scopus/ Web of Science

Изчислителната биология и молекулярната биология са области, които променят знанията и уменията за придобиване, съхраняване, управление, анализ, интерпретация и разпространение на биологична информация. Това изисква използването на високопроизводителни компютри и иновативни софтуерни инструменти за управление на огромна информация, както и внедряване на иновативни алгоритмични техники за анализ, интерпретация и прогнозиране на данни. Подравняването на последователности е важен метод в анализа на ДНК и протеини. Публикацията описва изчислителните предизвикателства при обработката на биологични последователности. Големите предизвикателства са да се предложат паралелни изчислителни модели и паралелни програмни реализации, базирани на алгоритмите за подравняване на биологични последователности. Представено е изследване на ефективността на подравняването на последователности, базирано на паралелна многонишкова програмна реализация на алгоритъма на

Needleman-Wunsch. Проектиран е паралелен изчислителен модел, базиран на алгоритъма на Needleman-Wunsch. Предложеният паралелен модел е верифициран чрез многонишкова паралелна програмна имплементация, използваща OpenMP на 8-ядрен сървър Xeon. Проведени са редица експерименти за различни набори от данни и различен брой нишки. Времето за изпълнение и ускоряването на параметрите на паралелната производителност са оценени експериментално. Оценката на производителността и анализите на мащабируемостта показват, че предложеният модел има добра мащабируемост както по отношение на натоварването, така и на размера на машината, като се мащабира по-добре с увеличаване на броя на ядрата.

Г.7.15. **Gancheva V.** A Big Data Management Approach for Computer Aided Breast Cancer Diagnostic System Supporting Precision Medicine, *AIP Conference Proceedings 2172*, 090012(2019), International Conference on Application of Mathematics in Engineering and Economics (AMEE'19), AIP Conference Proceedings, Scopus/ Web of Science, SJR = 0.190 (2019)

Представен е подход за управление на голямо количество разнородни набори от данни от различни източници на данни за компютърно подпомагана диагностична система за рак на гърдата. Архитектурата на големи геномни данни се състои от източници на данни, съхранение, интеграция и предварителна обработка, реален поток от данни, обработка на потоци, съхранение на аналитични данни, анализ и докладване. Дефинирани са дейностите по управление на данни за диагностична система за рак на гърдата. Проектирана е концептуална архитектура на база данни за съхраняване на набори от данни от няколко типа, за да се подпомогне прогнозирането на рак на гърдата. Базата данни за рак на гърдата се състои от информация, свързана с гени и функции на рак на гърдата - идентификатор, име, тип, организъм, функция и кодирани протеини, описание, връзка за извличане на последователност. Базата данни на пациента се състои от индивидуални данни за пациента - генетични данни, клинична история, параметри на индивидуалния начин на живот, резултати от клинични тестове, фактори на околната среда. Наборите от данни в предложената система за управление на големи данни се извличат от базите данни за биомедицински изследвания. Системата за управление на данни е независима от платформата, лесна за използване и осигурява достъп до други бази данни като PubMed, NCBI. Целта е да се използва за съхранение на данни в система за анализ на големи данни и откриване на знания, особено за казус за диагностика на рак на гърдата. Предимствата в управлението на данни, анализа и откриването на знания дават възможност на учените да постигнат нови научни пробиви. В резултат на това изследователската работа е насочена към бързо управление и обработка на клинични данни за решаване на проблеми в областта на прецизната медицина.

Г.7.16. **Gancheva V.** SOA Based Multi-Agent Approach for Biological Data Searching and Integration, *International Journal of Biology and Biomedical Engineering*, ISSN: 1998-4510, Volume 13, 2019, pp. 32-37, Scopus, SJR = 0.191 (2019), Q4

Основно предизвикателство при анализа на биологични данни е да се предложи интегриран и модерен достъп до прогресивно нарастващите количества данни в множество формати и ефективни подходи за тяхната обработка. Статията представя изследване на моделите за съхранение, извличане и интегриране на голямо количество геномни данни, както и решения за проблеми, свързани с хетерогенността, разпространението и съвместимостта на данните. Предлага се базиран на архитектура ориентирана към услуги (SOA) многоагентен подход за търсене и интегриране на биологични данни. Проектирана е концептуална архитектура за интегриране на разпределени биологични данни, базирана на SOA. Архитектурата е насочена към автоматизиране на интеграцията

на данни и позволява бързо управление на големи обеми от разнообразни набори от данни, представени в различни формати - релационни, NoSQL, плоски файлове. Интегрирането на различни бази данни се решава чрез използване на мултиагентна архитектура. Интеграционната система се състои от услуги за трансформиране на общата заявка в заявка на специфичен език за всяка локална база данни, в зависимост от нейния тип. Концептуалната интеграция на базата данни е решена чрез прилагане на подхода на транслиращи заявки. Всяка интегрирана база данни е представена от отделна концептуална схема, наречена виртуална схема. Тази схема се генерира в процеса на съпоставяне, който сравнява структурни елементи от базата данни с концептуалния модел. Мултиагентната система е ориентирана към услуги за търсене на биологични данни от различни източници, която изпраща заявки до множество бази данни и след това компилира резултатите в списък, в зависимост от вида на разработения източник. Системата позволява на потребителя да задава критерии за търсене и достъп до множество бази данни едновременно. Услугите позволяват системата да бъде достъпна през Интернет от множество клиенти (мобилни телефони, уеб браузъри, настолни приложения) и да обслужва широк кръг потребители едновременно.

Г.7.17. Borovska P., **Gancheva V.**, Ko S.-H., Scaling of Parallel Multiple Sequence Alignment on the Supercomputer JUQUEEN, Proceedings of the International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications IDAACS'2013, Berlin, German, ISBN 978-1-4799-1426-5, pp. 687-691, DOI: 10.1109/IDAACS.2013.6663013, Scopus

В публикацията се предлага оптимизация, мащабиране, оценка на производителността и профилиране на паралелно подравняване на множество последователности, базирано на алгоритъма ClustalW на суперкомпютъра BlueGene/Q, така наречения JUQUEEN, за казуса на последователностите на грипния вирус. За тази цел е проектиран и верифициран паралелен I/O интерфейс за едновременен и независим колективен достъп до един файл, на базата на паралелна програмна реализация на суперкомпютъра JUQUEEN. Направена е оценка на паралелната производителност и профилиране на множествено подравняване на базата на паралелна програмна реализация на алгоритъм ClustalW, използващ интерфейс за предаване на съобщения на суперкомпютъра JUQUEEN. Паралелни параметри на производителност като време за изпълнение, мащабиране и профилиране са оценени експериментално. Анализите на оценката на производителността и профилирането показват че паралелната система е добре балансирана както по отношение на работното натоварване, така и по отношение на размера на машината, с изключение на процеса с ранг 0, който е най-силно използван поради ефективността на комуникацията и синхронизацията. Паралелната I/O реализация води до силно ускоряване на първоначалната процедура и достига до 6,8 пъти по-бърз от базовия код в случай на 8K ядра на JUQUEEN. И все пак цялостната печалба от паралелното четене не е силна, тъй като размерът на входния файл е ограничен поради размера на разпределената памет. Разпределението на паметта на JUQUEEN за ядро е 1 GB, което ограничава максималния брой входни последователности до приблизително 10 000 (в зависимост от дължината на последователностите).

Г.8.1. **Gancheva V.** Intelligent Management and Analytics of Big Biomedical Scientific Research, Proceedings of XV International Scientific Conference e-Governance and e-Communications, Sozopol, June 2023

Статията представя предизвикателствата при създаване на интегрирана отворена технологична платформа за прилагане на интелигентни решения за управление и анализ на многомерни големи биомедицински данни, автоматизираща ефективни методи и алгоритми за анализ на големи биомедицински данни и прилагаща модели за тяхната визуализация. Основната цел на платформата

е да подпомогне извличането на знания и взимането на решения за нуждите на медицината и биологията, като предложи интегрирано решение за управление, съхранение, анализ и визуализация на големи масиви от разнородни данни и предостави лесна за използване инфраструктура за провеждане на научни изследвания и повишаване на ефективността. Характеристиките на платформата ще бъдат валидирани и демонстрирани чрез скалируема работна рамка с реконфигурация на ресурсите, като за целта ще се моделират различни научни работни потоци за анализ на медицински изображения и анализ на ДНК секвенции.

Г.8.2. Todorova V., **Gancheva V.**, Mladenov V. COVID-19 Medical Data Integration Approach, *Journal Molecular Sciences and Applications*, Volume 2, pp. 102-106, ISSN / EISSN : 2944-9138 / 2732-9992, DOI: 10.37394/232023.2022.2.11

Предложен е концептуален модел за интегриране и обработка на медицински данни, състоящ се от три слоя и шест фази. Управлението на данни се състои от три основни фази: подготовка на данните за анализ, интерпретация и визуализация, а подготвителната фаза включва събиране, съхранение, интегриране на медицински данни. Вторият слой за анализ на данни включва прилагане на методи за обработка на медицински данни. Процесът на обработка на данни включва манипулиране на събраните данни и извършване на функции и операции с цел извличане на значима информация като валидиране, сортиране, обобщаване, анализ, докладване, класификация. Фазата на класифициране на медицински данни включва процеса на подравняване на данни в групи въз основа на предварително определени критерии. За целите на данните се прилагат методи и техники за клъстериране като k-Nearest Neighbor (kNN), kMeans, Support Vector machine (SVM), Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Naive Bayes и др. Проектиран е работен процес за интегриране на медицински данни, включително стъпки за интегриране, филтриране, агрегиране и сортиране на данни. Предложеният работен процес е валидиран за медицински данни за SARS-CoV-2 от клинични досиета на 20400 потенциални пациенти.

Г.8.3. Borovska P., **Gancheva V.** Massively Parallel Multiple Sequence Alignment on the Supercomputer JUQUEEN, *NAUN International Journal of Computers*, Vol. 12, 2018, pp. 1-8, ISSN: 1998-4308

Обработката на биологичните последователности е ключова задача в молекулярната биология. Тази научна област изисква мощни изчислителни ресурси за изследване на големи набори от биологични данни. Паралелните in silico симулации, базирани на методи и алгоритми за анализ на биологични данни с помощта на високопроизводителни разпределени изчисления, са от съществено значение за ускоряване на изследванията и намаляване на инвестициите. Подравняването на множество последователности е широко използван метод за обработка на биологични данни. Публикацията е фокусирана върху изследване на производителността и подобряване на софтуера за подравняване на множество биологични последователности MSA_BG на суперкомпютъра BlueGene/Q JUQUEEN. Експериментални симулации на базата на паралелна имплементация на алгоритъма MSA_BG за подравняване на множество последователности са проведени за казуса на изследването на вариабилността на грипния вирус. Целите на изследването са оптимизиране на кода, мащабиране, профилиране и оценка на производителността на софтуера MSA_BG. Разработено е хибридно MPI/OpenMP паралелизиране и са показани предимствата на този подход чрез резултатите от бенчмарк тестове, извършени на JUQUEEN. Експерименталните резултати показват, че хибридната паралелна реализация осигурява значително по-добра производителност от реализация само на MPI.

Г.8.4. Ivanova D., Borovska P., **Gancheva V.** Experimental Investigation of Enhancer-Promoter Interactions out of Genomic Big Data based on Machine Learning, *International Journal of Computers*, Volume 3, 2018, ISSN: 2367-8895, pp. 58-62

Основната цел на статията е да представи експериментално изследване за откриване на взаимодействия между енхансер-промотор от геномни големи данни, базирано на машинно обучение, предлагащ работен процес за откриване на енхансер-промотор взаимодействия. Реализира се чрез използване на класификаторите Decision Tree и Support Vector Machine. Експерименталната рамка е базирана на средата на Apache Spark, която позволява поточно предаване и анализ на големи данни в реално време. Библиотеката за машинно обучение на Apache Spark (MLlib) е имплементирана на език за програмиране Python за обработка на големи геномни данни. За постигане на резултатите са използвани данни за взаимодействията между енхансер-промотор GM12878 и K562. Представени и обсъдени са получените експериментални резултати.

Г.8.5. Borovska P., **Gancheva V.** Parallelization and Optimization of Multiple Biological Sequence Alignment Software Based on Social Behavior Model, *International Journal of Computers*, pp. 69-74, ISSN: 2367-8895, Volume 3, 2018,

Огромното количество биологични последователности, натрупани в световните бази данни за нуклеотиди и протеини, води до необходимостта от ефективни инструменти за структурен геномен и функционален анализ. Тази научна област изисква мощни изчислителни ресурси за изследване на големи набори от биологични данни. Подравняването на множество последователности е важен метод в ДНК и протеиновия анализ и обикновено представлява подравняване на три или повече биологични последователности с подобна дължина. В резултат на обработката може да се извлече хомология и да се изследват еволюционните връзки между последователностите. Целта на тази публикация е да предложи паралелизиране и оптимизиране на софтуера за подравняване на множество последователности MSA_BG, за да се подобри производителността, за казуса на последователностите на грипния вирус. Целта е оптимизиране на кода, пренасяне, мащабиране и оценка на производителността на паралелния софтуер за подравняване на множество последователности MSA_BG за Intel Xeon Phi (архитектура MIC). За тази цел е внедрена и проверена паралелна многонишкова оптимизация, включително OpenMP. Експерименталните резултати показват, че хибридната паралелна реализация, използваща MPI и OpenMP, осигурява значително по-добра производителност от оригиналния код.

Г.8.6. Borovska P., **Gancheva V.**, Georgiev I. Hybrid Parallel Implementation of Multiple Sequence Alignment Software ClustalW on Intel Xeon Phi, Proceeding of Sixth International Conference on Advances in Computing, Electronics and Communication - ACEC 2017, Rome, Italy, Page(s) : 47 – 51, Electronic ISBN : 978-1-63248-138-2, DOI: 10.15224/ 978-1-63248-138-2-10

Представената разработка е насочена към изследване и подобряване на производителността на софтуера за подравняване на множество последователности ClustalW на тестовата платформа EURORA в CINECA, за казус на последователностите на грипния вирус. Целта е оптимизиране на кода, пренасяне, мащабиране и оценка на производителността на паралелния софтуер за подравняване на множество последователности ClustalW за Intel Xeon Phi (архитектурата MIC). За тази цел е внедрена и проверена паралелна многонишкова оптимизация, включително OpenMP. Експерименталните резултати показват, че хибридната паралелна реализация, използваща MPI и OpenMP, осигурява значително по-добра производителност от оригиналния код.

Г.8.7. Ташев Т., Лазарова М., **Ганчева В.**, Иванова В. Обучение по управление на ресурси в предприятията, Национална конференция с международно участие "ОБРАЗОВАТЕЛНИ ТЕХНОЛОГИИ 2014", Каварна, 12-14 септември 2014

Предприятията в целия свят прилагат система за планиране и управление на ресурсрите (Enterprise Resource Planning - ERP). Това налага обучението на студенти в магистърска степен в областта в доста университет в Ервопа и САЩ. Направено е проучване и детайлен анализ на аналогични специалности за бакалавърско и магистърско обучение в чуждестранни университети, с цел да се изяснят добрите практики при обучението на студентите. След обстоен анализ и определяне на необходимите компетенции, знания и умения, както и потребността на бизнеса от специалисти в областта, е създаден учебен план към Факултета за Английско инженерно обучение. Разработени са иновативни дисциплини в резултат на обучението, по които се осигуряват специалисти по управление на ресурси в предприятията, които да отговарят на нуждите на пазара на труда в България и ЕС.

Г.8.8. Borovska P., Dokomes H., **Gancheva V.**, Tsvetanov S. GRID Resource Broker Architecture Based on Metadata Scheduling Model, *Journal Computer & Communications Engineering*, Vol. 7 No 1/2013, pp. 5-12, ISSN 1314-2291

В статията се предлага архитектура на GRID ресурсен брокер, базиран на модел за планиране с метаданни. Предложен е сценарий за планирането, при който моделът за планиране обхваща 5 базови модела: на ресурсите, приложението, производителността, метрика за производителността, политика за планиране и програмен модел. Дефинирани са множествата метаданни на входа и на изхода на модела на планиране. Предложеният архитектурен проект на GRID ресурсен брокер включва три базови модула: картограф, оценител цена/ресурси и диспечер. Предложен е сценарий за ресурсен мениджмънт в GRID, като са дефинирани взаимодействията на основните модули на ресурсния брокер с услугите в GRID.

Г.8.9. Borovska P., **Gancheva V.**, Aleksieva-Petrova A., Dokomes H. Virtual Center for In-Silico Science and Technology Transfer ViSTa, *Journal Computer & Communications Engineering*, Vol. 7 No 2/2013, pp. 5-11, ISSN 1314-2291

Целта на виртуалния център за in-silico трансфер на наука и технологии е създаването на виртуална платформа за е-наука, която предоставя електронни ресурси от наука, експертиза, компютърни модели и софтуерни инструменти за биосимулации и биологична база данни за молекулярна биология, геномика, виртуален скрининг за дизайн на лекарства и свързани области на науките за живота. Инфраструктурата осигурява богата функционалност от инструменти и услуги, позволява динамично свързване на изчислителни ресурси, реализиране на разпределени изчисления и постигане на висока производителност. Виртуалната платформа предоставя възможност за изграждане на гъвкава мрежа от независими звена, свързани с информационните технологии за споделяне на умения, in-silico наука и технологии и предоставяне на достъп до нетрадиционен опит на други звена. Виртуалната платформа осигурява удобен за потребителя интерфейс, който улеснява максимално използване от биолози, химици, генетици и други. Използват се иновативни инструменти като научни портали, агент-базирани технологии и онтологии. Създадена е работна рамка за интегриране на GRID и CLOUD услуги, така че да предоставят на приложенията единен достъп до данни и ресурси.

Г.8.10. Боровска П., **Ганчева В.**, Цветанов С., Оптимизация и изследване на паралелната производителност на софтуерен пакет GADGET на суперкомпютър BlueGene/P и паралелни системи с GPGPU ускорители, сп. *Автоматика и информатика*, год. XLVII, 2/2013, стр. 19-27, ISSN 0861-7562

В статията е представен подход за оптимизация на софтуерния пакет Gadget. На базата на статичен и динамичен анализ на програмата Gadget са установени слабите места и са определени фрагментите, които са видифицирани и съответно изпълнявани на специализирани хибридни архитектури, използващи ускорители с цел повишаване на производителността, като е показан и илюстриран пример за конкретна имплементация. Направена е оценка на паралелната производителност – времена на изпълнение, скалиране, профилиране, на базата на експериментални симулации на суперкомпютър BlueGene/P и GPGPU паралелни системи. Експерименталните резултати показват, че след направената цялостна оптимизация на програмния код, ускорението се увеличава приблизително с 50% при добра скалируемост.

Г.8.11. Borovska P., **Gancheva V.**, Landzhev N. High Performance Grid Environment for Parallel Multiple Biological Sequence Alignment, Proceedings of The Eighth International Multi-Conference on Computing in the Global Information Technology ICCGI'2013, Nice, France, ISBN 978-1-61208-283-7, pp. 82-87

Паралелните in silico симулации, базирани на методи и алгоритми за анализ на биологични данни с помощта на високопроизводителни разпределени изчисления, са от съществено значение за ускоряване на изследванията и намаляване на инвестициите. Тази публикация представя високопроизводителна Grid среда, интегрираща различни услуги и софтуер за улесняване на достъпа до разпределени ресурси за провеждане на научни експерименти в областта на биоинформатиката. Средата позволява паралелни компютърни симулации, повишавайки ефективността на изчисленията и позволявайки на учените лесен и удобен достъп. Уеб порталът предоставя като услуги достъп и изпълнение на паралелна програмна реализация на базата на алгоритъм за сравнителен анализ на биологични данни. Предложеният портал е верифициран експериментално за казус за изследване на изменчивостта на грипния вирус. Проектиран е иновативен паралелен алгоритъм MSA_BG за множествено подравняване на биологични последователности, който е силно мащабируем. Алгоритъмът MSA_BG е итеративен и се основава на концепцията за метаевристика на изкуствена пчелна колония и концепцията за корелация на алгоритмични и архитектурни пространства. Конструирана е метафората на метаевристиката ABC и са дефинирани функционалностите на агентите. Проектиран е концептуалният паралелен изчислителен модел. Конструирана е алгоритмичната рамка на проектирания паралелен алгоритъм. Алгоритъмът MSA_BG има йерархична структура, която позволява спазване на принципа на локалност (независими изчисления) и много висока скалируемост (кошери и рояци), така че се очакват високоефективни реализации за петафлопс суперкомпютри. Направена е оценка на паралелната производителност и профилиране на подравняване на множество последователности на базата на алгоритъм MSA_BG, хетерогенни разпределени високопроизводителни изчислителни ресурси. Казусът изследва нуклеотидните последователности на грипния вирус и открива консенсусни мотиви и променливи домейни в различните сегменти. Паралелните параметри на производителност, като време за изпълнение и ускорение, са оценени експериментално. Анализите на оценката на производителността показват, че паралелната система е добре балансирана както по отношение на натоварването, така и по отношение на размера на машината.

Г.8.12. Borovska P., **Gancheva V.**, Georgiev I. Optimization of Multiple Sequence Alignment Algorithm ClustalW Using OpenMP and Vector Processing, Proceedings of the 11th International Conference on Challenges in Higher Education and Research in the 21st Century, 2013, Sozopol, Bulgaria, ISBN 978-954-580-325-3, pp. 181-184, https://elfe.tu-sofia.bg/cher21/index.php?nact=5021

Биоизчисления и молекулярна биология са области, изискващи знания и умения за придобиване, съхраняване, управление, анализ, интерпретация и разпространение на биологична информация. Това изисква използването на високопроизводителни компютри и иновативни софтуерни инструменти за управление на огромна информация, както и внедряването на иновативни алгоритмични техники за анализ и интерпретация на данни. В тази статия е направено изследване и сравнителни анализи на нуклеотидни последователности на грипния вирус на базата на паралелна компютърна симулация. За тази цел е предложен паралелен многонишков изчислителен модел, базиран на алгоритъм ClustalW за множествено подравняване на последователности. Предложеният модел е верифициран на базата на изпълнение на паралелна програмна имплементация върху разнороден компактен компютърен клъстер.

Г.8.13. Borovska P., **Gancheva V.**, Asenov E., Georgiev I. Computational Aspects of In-silico Experiments for Investigating the Impact of the Host Genome on the Influenza Virus A Variability, *Journal Information Technologies and Control*, Vol.10, No 2/2012, pp. 8-14, ISSN 1312-2622, http://www.acad.bg/rismim/itc/sub/archiv/no2_2012.htm

В днешно време изследването на промените на грипния вирус е проблем с изключително голямо значение. Вирусите на грип тип А причиняват епидемии и пандемии. Проблемът за ограничаване на разпространението на пандемии и лечението на хората, заразени с грипния вирус, се основава до голяма степен на най-новите постижения на молекулярната биология, биоинформатиката, както и много други напреднали области на науката. Обработката на биологични последователности in silico е ключ за молекулярната биология. Тази научна област изисква мощни изчислителни ресурси за изследване на големи набори от биологични данни. Статията представя паралелни изчислителни симулации за казуса изследване на ролята на генома на гостоприемника в еволюцията и бързата променливост на грипния вирус А на суперкомпютър BlueGene/P. Експерименталната рамка се основава на всички налични съществуващи нуклеотидни последователности на грипния вирус А, алгоритъма ClustalW за подравняване на множество последователности, алгоритъма Blast за търсене на последователности, софтуера Philip за реконструкция на филогенетично дърво и инструмента за рекомбинационен анализ за намиране на горещи точки на мутация/ рекомбинация в геномите на вируса на грип А.

Г.8.14. Borovska P., **Gancheva V.**, Tsvetanov S. Optimization and Scaling of Multiple Sequence Alignment Software ClustalW on Intel Xeon Phi, PRACE White Paper, 2014, Available online at https://prace-ri.eu/wp-content/uploads/wp138.pdf

Тази работа е насочена към изследване и подобряване на производителността на софтуера за подравняване на множество последователности ClustalW на тестовата платформа EURORA в CINECA, за казус на последователностите на грипния вирус. Целта е оптимизиране на кода, пренасяне, мащабиране и оценка на производителността на паралелния софтуер за подравняване на множество последователности ClustalW за Intel Xeon Phi (архитектура MIC). За тази цел е внедрена и проверена паралелна многонишкова оптимизация, включително OpenMP. Експерименталните

резултати показват, че хибридната паралелна реализация, използваща MPI и OpenMP, осигурява значително по-добра производителност от оригиналния код.

Г.8.15. Borovska P., **Gancheva V.**, Landzhev N. Code Optimization and Scalability Testing of an Artificial Bee Colony Based Software for Massively Parallel Multiple Sequence Alignment on the Intel MIC Architecture, PRACE White Paper, 2014, Available online at https://prace-ri.eu/wp-content/uploads/wp137.pdf

Тази дейност от проекта PRACE има за цел да проучи и подобри производителността на софтуера за подравняване на множествени последователности MSA_BG на компютърната система EURORA в CINECA, за казус на последователностите на грипния вирус. Целта е оптимизиране на кода, пренасяне, мащабиране и оценка на производителността на паралелния софтуер за подравняване на множество последователности MSA_BG за Intel Xeon Phi (архитектура MIC). За тази цел е внедрена и верифицирана паралелна многонишкова оптимизация, вкл. OpenMP. Експерименталните резултати показват, че хибридната паралелна реализация, използваща MPI и OpenMP, осигурява значително по-добра производителност от оригиналния код.

Г.8.16. Charalampidou A., Daoglou P., Folias D., Borovska P., **Gancheva V.** A Hybrid Implementation of Massively Parallel Multiple Sequence Alignment Method Based on Artificial Bee Colony Algorithm, PRACE White Paper, 2014, Available online at https://prace-ri.eu/wp-content/uploads/wp127.pdf

Проектът се фокусира върху изследване на производителността и подобряване на софтуера за подравняване на множество биологични последователности MSA_BG на суперкомпютъра BlueGene/Q JUQUEEN. За тази цел са проведени научни експерименти в областта на биоинформатиката, като са използвани като казус последователности на грипния вирус. Целите на проекта са оптимизиране на кода, настройка, мащабиране, профилиране и оценка на производителността на софтуера MSA_BG. За тази цел е разработена хибридна MPI/OpenMP паралелизация в горната част на кода само за MPI и са демонстрирани предимствата на този подход чрез резултатите от сравнителни тестове, извършени на JUQUEEN. Експерименталните резултати показват, че хибридната паралелна реализация осигурява значително по-добра производителност от оригиналния код.

Г.8.17. Borovska P., **Gancheva V.** Massively Parallel Algorithm for Multiple Sequence Alignment Based on Artificial Bee Colony, PRACE White Paper, 2013, Available online at https://prace-ri.eu/wp-content/uploads/wp114.pdf

Тази дейност с проекта PRACE-2IP има за цел да проучи и подобри производителността на софтуера за подравняване на множество последователности ClustalW на суперкомпютъра BlueGene/Q, така наречения JUQUEEN, за казус на последователностите на вируса на грипа. Пренасянето, настройката, профилирането и мащабирането на този код са извършени в този аспект. Проектиран е паралелен входно-изходен интерфейс за ефективно последователно въвеждане на набор от данни, в който локалните главни на подгрупите се грижат за операцията по четене и излъчват набора от данни към своите подчинени устройства. Оптималният размер на групата е изследван и ефектите от размера на буфера за четене върху производителността на четене са експериментирани. Приложението към софтуера ClustalW показва, че текущата реализация с паралелен I/O осигурява значително по-добра производителност от оригиналния код с оглед на I/O сегмента, което води до 6,8 пъти ускорение за въвеждане на набор от данни в случай на използване на 8192 JUQUEEN ядра.

Г.8.18. Ko S.-H., Borovska P., **Gancheva V.** Optimization of Multiple Sequence Alignment Software ClustalW, PRACE White Paper, 2013, Available online at https://prace-ri.eu/wp-content/uploads/wp71.pdf

Обработката на биологичните последователности е ключова задача в молекулярната биология. Тази научна област изисква мощни изчислителни ресурси за изследване на големи набори от биологични данни. Паралелните in silico симулации, базирани на методи и алгоритми за анализ на биологични данни с помощта на високопроизводителни разпределени изчисления, са от съществено значение за ускоряване на изследванията и намаляване на инвестициите. Подравняването на множество последователности е широко използван метод за обработка на биологични последователности. Целта на този метод е подравняване на ДНК и протеинови последователности. Тази статия представя иновативен паралелен алгоритъм MSA_BG за множествено подравняване на биологични последователности, който е силно мащабируем и съобразен с местоположението. Алгоритъмът MSA_BG е итеративен и се основава на концепцията за метаевристика на изкуствена пчелна колония и концепцията за корелация на алгоритмични и архитектурни пространства. Конструирана е метафората на метаевристиката ABC и са дефинирани функционалностите на агентите. Проектиран е концептуалният паралелен модел на изчисление и е конструирана алгоритмичната рамка на проектирания паралелен алгоритъм. Експериментални симулации на базата на паралелно внедряване на алгоритъм MSA_BG за подравняване на множество последователности върху хетерогенен компактен компютърен клъстер и суперкомпютър BlueGene/P са проведени за казуса от изследването на вариабилността на грипния вирус. Анализите на оценката на производителността и профилирането показват, че паралелната система е добре балансирана както по отношение на работното натоварване, така и по отношение на размера на машината.

Публикации с IF/SJR

3.31.1.Sharabov M., Tsochev G., **Gancheva V.**, Tasheva A. Filtering and Detection of Real-Time Spam Mail Based on a Bayesian Approach in University Networks. *Electronics*. 2024; 13(2):374. . **IF=2.9 (2022) / SJR=0.644 (2023) / Q2, Scopus / WoS**

С навлизането на цифровите технологии като неразделна част от днешното ежедневие, рискът от пробиви в информационната сигурност нараства. Спамът по имейл, известен като нежелана поща, продължава да представлява значително предизвикателство в дигиталната сфера, заливайки входящи кутии с нежелани и често неподходящи съобщения. Този безмилостен приток на нежелана поща не само нарушава продуктивността на потребителите, но също така поражда опасения за сигурността, тъй като често служи като средство за опити за фишинг, разпространение на злонамерен софтуер и други киберзаплахи. Разпространението на спама се подхранва от евтиното му разпространение и способността му да достига до широка аудитория, използвайки уязвимости в системите за електронна поща. Публкацията бележи началото на задълбочено проучване на жизнеспособността и потенциалното внедряване на стабилна система за филтриране и предотвратяване на спам, специално пригодена за университетските мрежи. С ескалиращата заплаха от хакерски атаки, базирани на електронна поща, и непрекъснатия поток от спам, необходимостта от всеобхватен и ефективен механизъм за защита в рамките на академичните институции става все по-наложителна. Проучвайки потенциални решения, това изследване се задълбочава в приложимостта и ефикасността на Bayes филтри, клас вероятностни класификатори, известни със своята способност да разграничават легитимни имейли от спам съобщения. Bayes

филтрите използват статистически алгоритми за анализиране на съдържанието на имейлите, моделите на обучение и функциите за точно категоризиране на входящите имейли. Резултатите, получени от подхода на Bayes са положителни, макар и да не достигат задоволителни нива. Очевидно е, че съществените подобрения могат значително да повишат ефикасността на предложения модул за филтриране на спам чрез внедряване на различни стратегии. В обобщение, експерименталните открития потвърждават уместността на графиките на Bayes в сферата на филтрирането на спам. Очевидно е обаче, че са наложителни съществени подобрения. Като такива, бъдещите усилия ще се съсредоточат върху включването на нови прозрения за данни и стриктното оценяване на ефективността на алтернативните класификатори. Тези колективни усилия са готови да осигурят окончателно решение, което надхвърля настоящите ограничения, създавайки стабилен и адаптивен механизъм за филтриране на нежелана поща за оптимална сигурност на електронната поща.

3.31.2. Gancheva V., Galabova L. Platform for Learning and Virtual Reality in Animal Husbandry, *WSEAS Transactions on Information Science and Applications*, pp. 163-169, 2023, DOI 10.37394/23209.2023.20.19 Scopus, SJR = 0.126 (2023), Q4

В днешно време цифровите технологии се използват широко в сферата на образованието. Виртуалната и добавена реалност и 3D технологиите навлизат в сферата на образованието на всички образователни нива. Те са предпоставка за прилагане на нови подходи при представяне на учебното съдържание и по-лесното му възприемане и усвояване от обучаемите. Работата, представена в публикацията, е насочена към изследване и анализ на системи, методи и инструменти за дигитализация на образованието и създаване на нови образователни ресурси в областта на животновъдството като инструменти и възможности за създаване на нови образователни ресурси в областта на животновъдството, базирано на разширена и виртуална реалност, и използване на триизмерни (3D) модели за визуализиране на учебно съдържание. В публикацията е представена интегрирана платформа за отворена наука и споделяне на образователни ресурси, както и среда за дистанционно обучение и анализ на данни в животновъдството, извлечени от учебните ресурси на даден курс в системата. Изследването е насочено към разработване на компютърно подпомагана рамка в областта на дигитализираното образование и създаване на нови образователни ресурси за дистанционно обучение в животновъдството. Предложената платформа предоставя различни начини за достъп и споделяне на образователни ресурси чрез цифрови технологии и хранилище за предоставяне на достъп до безплатни онлайн курсове и учебно съдържание.

3.31.3. Gancheva V., Georgiev I., Todorova V. X-Ray Images Analytics Algorithm based on Machine Learning, WSEAS Transactions on Information Science and Applications, 2023, pp. 136-145, DOI: 10.37394/23209.2023.20.16, Scopus, SJR = 0.126 (2023), Q4

Бързото развитие на информационните технологии доведе до огромно количество данни, генерирани от големи или сложни системи и устройства. Приложенията в информационните технологии, медицината и много други области генерират големи обеми данни, които предизвикват анализаторите на данни. Анализът на данни намира приложение в области, където статистическите и аналитични методи и изградените чрез тях модели не са достатъчни. В публикацията се обсъждат източниците на медицински данни, случаи на употреба и анализ на данни в медицината, както и методи и алгоритми за анализ на данни. Целта и задачите на изследването, представени в публикацията, са да се предложи алгоритъм за обработка на рентгенови изображения, базиран на инструменти и техники от областта на машинното обучение. Фазата на предварителна обработка

включва трансформация на изображенията, извличане на функции и избор на набори от данни за обучение и тестване. Предварителната обработка на данни дава възможност за обработка на данни, които иначе не биха били подходящи, чрез коригиране на данните към спецификациите, установени от всяка процедура за извличане на данни. Всяка характеристика се изследва на втория етап, за да се идентифицират и класифицират всички потенциални модели. В последния етап с помощта на алгоритъм за машинно обучение се избира най-ефективният модел за извличане на шаблон на модела или поведение на данните. Предложеният алгоритъм е тестван с помощта на публично достъпни набори от данни за рентгенови изображения с отворен код за обучение и тестване на предложения подход, състоящи се от четири класа: нормално, белодробна непрозрачност, пневмония и COVID-19. За целите на системното тестване и валидиране на алгоритъма е проектиран работен процес за класификация на медицински изображения. В експерименталния работен процес са определени и внедрени пет алгоритъма в областта на машинното обучение: логистична регресия, Naive Bayes, Random Forest, SVM и невронна мрежа. В сравнение с резултатите от Random Forest, Logistic Regression, Naive Bayes и SVM, констатациите от експерименталния анализ и резултатите показват, че невронните мрежи дават най-добри резултати и тези резултати могат да се считат за най-надеждни.

3.31.4. Aleksieva-Petrova A., Gancheva V., Petrov M. APTITUDE Framework for Learning Data Classification Based on Machine Learning, *International Journal of Circuits, Systems and Signal Processing*, Volume 14, 2020, https://doi.org/10.46300/9106.2020.14.51, SJR=0.156 (2020) Q4, Scopus

Анализът на обучението се свързва с прилагане на машинно обучение за осигуряване на прогнози за успеха на обучаемите и предписания за обучаеми и учители. Основната цел на публикацията е да предложи рамка APTITUDE за класифициране на учебни данни, за да се постигне адаптиране и препоръки на съдържанието на курса или потока от дейности на курса. Тази рамка прилага модел за прогнозиране на обучението на студенти, базирано на машинно обучение. Пет алгоритма за машинно обучение се използват за осигуряване на класификация на учебни данни: random forest, Naïve Bayes, k-nearest neighbors, logistic regression и support vector machines. Предложената рамка APTITUDE помага за структуриране и съхранение на големи данни от хетерогенни източници както като LMS, така и като образователна игра; идентифициране на модели, като анализира поведението на обучаемите и позволява анализи на данни с описателни, предсказващи и предписващи резултати. Проектиран е алгоритъм за прогнозиране на обучението на студентите, базиран на машинно обучение за обработка и анализ на данни и откриване на знания по отношение на основните дейности на обучаемия и учителя. Експерименталният набор от данни е получен от системата за управление на обучението и съдържа 63774 екземпляра, характеризиращи се със 7 атрибута. Използвани са лог файлове в системата Moodle за анализи. За анализ на данните се използват алгоритми за групиране, предоставени: максимизиране на очакванията, йерархично групиране, прости K-средни стойности и X-средни стойности за намиране на корелация в степенувани дейности. Имплементирани са други пет алгоритъма от областта на машинното обучение, за да се потвърди тяхната приложимост за прогнозиране на обучението на студентите. Въз основа на различни аналитични модели, които са създадени след изпълнението на процеса на извличане на функции и намаляване на набор от данни, прототипът ще бъде валидиран и ще провери използваемостта на предложената архитектура.

SUMMARY OF SCIENTIFIC PAPERS

of assoc. prof. PhD Veska Stefanova Gancheva

submitted for participation in a competition for the position of PROFESSOR
professional field: 5.3. Communication and computer technology
scientific specialty: Systems with artificial intelligence
to the Department of Programming and Computer Technologies
Faculty of Computer Systems and technologies, Technical Univrsity of Sofia
published in SG no. 28 02.04.2024

In this competition are submitted scientific publications different from those included in the PhD thesis, the procedure and registration in the NACID register for academic position "Associate Professor" as follows: 50 scientific publications, incl. 32 scientific publications in publications that are referenced and indexed in Web of Science / Scopus; 2 textbooks and 1 study manual; 69 citations; 25 research and educational projects; supervision of 2 successfully defended PhD students, distributed by groups:

1. **Group B**
   1.1. Indicator B.4. Habilitation thesis - scientific publications, refereed and indexed (minimun 10) – 11.
2. **Group Г**
   2.1. Indicator Г.7. Scientific publication in refereed and indexed databases – 17.
   2.2. Indicator Г.8. Scientific publication in non-refereed peer-reviewed journals or in edited collective volumes – 18.
3. **Group Д**
   3.1. Indicator Д.12. Citations or reviews in refereed and indexed scientific publications – 58.
   3.2. Indicator Д.14. Citations in non-indexed peer-reviewed journals – 11.
4. **Group E**
   4.1. Indicator E.17. Supervision of a successfully defended PhD student – 2.
   4.2. Indicator E.18. Participation in national scientific or educational project – 14.
   4.3. Indicator E.19. Participation in international scientific or educational project – 8.
   4.4. Indicator E.20. Management of national scientific or educational project – 1.
   4.5. Indicator E.22. Funds raised for projects managed by the applicant – 1.
   4.6. Indicator E.23. Published university textbook or a textbook used on school network – 2.
   4.7. Indicator E.24. Published university manual or a manual used on school network – 1.
   4.8. Indicator E.29. Management of scientific or educational project – 2.
5. **Group Ж**
   5.1. Показател Ж.30. Schedule of lectures for the last three years – 722 hours.
6. **Group З**
   6.1. Indicator З.31. Научни публикации в списания с импакт фактор (IF на Web of Science) и/или с импакт ранг (SJR) на Scopus – 4.

Habilitation work – scientific publications (at least 10) published in editions that are referenced and indexed in world-recognized databases with scientific information

The submitted 11 scientific publications are unified by the main topic of "**Intelligent methods and tools for processing biomedical data**". The papers have been published in refereed and indexed in world-recognized scientific information databases Scopus/Web of Science after receiving the educational and scientific degree "doctor" and occupying the academic position "Associate Professor", in peer-reviewed scientific editions: international journals with IF/SJR (8) and proceedings of international conferences (3).

The scientific papers submitted for review address issues related to the use of artificial intelligence techniques, algorithms and tools for computer simulations in bioinformatics and medical image processing. The development of technologies for the generation of biomedical data - sequencers that generate genetic data and imaging tools - computed tomography images, nuclear magnetic resonance images, multi-layer microscopic images in cell analysis, etc. lead to the accumulation of a large volume of heterogeneous data. The past decade has seen an explosion in the amount of data available in bioinformatics and medicine. The amount of data is becoming so large that traditional data analysis platforms and methods can no longer meet the need to quickly perform data analysis and knowledge extraction tasks in the life sciences. Artificial intelligence and machine learning, including convolutional neural networks, are increasingly entering the fields of bioinformatics, healthcare and medicine. Bioinformatics is a rapidly developing field enabling scientific experiments through computer models and simulations. A challenge in data analysis is to offer integrated and modern access to the progressively growing volume of data, as well as efficient algorithms for their processing. The conducted research is related to the development of an integrated open technological platform for the development of work processes, consisting of a set of software tools for automating the computational process when conducting scientific experiments and implementing intelligent solutions for managing and extracting knowledge from multidimensional data. Methods and algorithms for the analysis of biomedical data, based on mathematical modeling and metadata synthesis, have been developed and optimized to ensure high quality of analysis, reduced computational complexity, and the possibility of parallel processing. The results are increased efficiency in the analysis of large arrays of biomedical data.

B.4.1. **Gancheva V.**, Stoev H. Optimization and Performance Analysis of CAT Method for DNA Sequence Similarity Searching and Alignment. *Genes*. 2024; 15(3):341, **IF=3.5 (2022) / SJR=0.817 (2023) / Q2, Scopus / WoS**

This paper presents a new version of the pairwise DNA sequences alignment algorithm, based on a new method called CAT, where a dependency with a previous match and the closest neighbor are taken into consideration to increase the uniqueness of the CAT profile and to reduce possible collisions, i.e., two or more sequence with the same CAT profiles. This makes the proposed algorithm suitable for finding the exact match of a concrete DNA sequence in a large set of DNA data faster. In order to enable the usage of the profiles as sequence metadata, CAT profiles are generated once prior to data uploading to the database. The proposed algorithm consists of two

main stages: CAT profile calculation depending on the chosen benchmark sequences and sequence comparison by using the calculated CAT profiles. Improvements in the generation of the CAT profiles are detailed and described in this paper. Block schemes, pseudo code tables, and figures were updated according to the proposed new version and experimental results. Experiments were carried out using the new version of the CAT method for DNA sequence alignment and different datasets. New experimental results regarding collisions, speed, and efficiency of the suggested new implementation are presented. Experiments related to the performance comparison with Needleman–Wunsch were re-executed with the new version of the algorithm to confirm that we have the same performance. A performance analysis of the proposed algorithm based on the CAT method against the Knuth–Morris–Pratt algorithm, which has a complexity of O(n) and is widely used for biological data searching, was performed. The impact of prior matching dependencies on uniqueness for generated CAT profiles is investigated. The experimental results from sequence alignment demonstrate that the proposed CAT method-based algorithm exhibits minimal deviation, which can be deemed negligible if such deviation is considered permissible in favor of enhanced performance. It should be noted that the performance of the CAT algorithm in terms of execution time remains stable, unaffected by the length of the analyzed sequences. Hence, the primary benefit of the suggested approach lies in its rapid processing capabilities in large-scale sequence alignment, a task that traditional exact algorithms would require significantly more time to perform. The approach of precomputing metadata and applying the trilateration principle provides a solution for the problem of slow alignment and similarity searching of biological data. Modification of the benchmark sequences and the way profiles are calculated and how they are compared results in the output of the comparison. This makes the approach adjustable to the desired level of accuracy. The experiments underscore the efficiency of the proposed algorithm and its potential to significantly speed up the process of DNA sequence alignment by leveraging the refined CAT profiles. The updated algorithm promises to be a valuable tool in bioinformatics, offering a faster and more reliable means for processing the vast and growing repositories of genetic data.

B.4.2. **Gancheva V.,** Stoev H., An Algorithm for Pairwise DNA Sequences Alignment. Bioinformatics and Biomedical Engineering. *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol 13919, 2023, **Scopus, SJR = 0.606 (2023), Q3**

A new algorithm for arranging DNA sequences based on the suggested CAT method is proposed, consisting of an algorithm for calculating a CAT profile against the selected reference sequences and an algorithm for comparing two sequences, based on the calculated CAT profiles. Implementation steps, inputs and outputs are defined. A software implementation of the proposed method for arranging biological sequences CAT has been designed and developed. Experiments have been carried out using different data sets to align DNA sequences based on CAT method. An analysis of the experimental results have been done in terms of collisions, speed and effectiveness.

The proposed new method for DNA sequences alignment, called CAT, based on the trilateration method, is experimentally verified. Three constant benchmarks have been established for the application of trilateration, which creates a constant favorite sequence - ie. independent of the records in the database and remains the same when the set is changed. Since the benchmark sequences established are constant (i.e. they do not depend either on the data or on their number), this allows the comparisons to be made at the very beginning – when the sequences is uploaded into database and this to be metadata information, accompanying each sequence. In

this way, there is no need to compare sequences during lookup (the slowest operation), but instead only the metadata generated when the data is entered is compared. The generation of the CAT profiles is done once during the data upload, which al-lows the profiles to be used as accompanying metadata information for the sequenc-es. Search comparisons with the CAT method are minimized and have a constant O(24) algorithm complexity, which helps optimize searches in large biological da-tasets and makes it suitable for implementation as a first step in more refined algo-rithms like FASTA. Based on the CAT profiles, sequences can be organized into a hierarchical storage structure to be used as a database for biological data storage in search-optimized systems. The paper presents the results of the developed program implementation of the pro- posed method CAT for biological sequences alignment. Experiments have been carried out with different datasets for DNA sequence alignment using the triplet-based CAT method. An analysis of the experimental results have been made. The analysis of the experimental results obtained by sequence alignment shows a small deviation of the proposed algorithm based on the CAT method, which can be ignored if this deviation is acceptable at the expense of performance. The execution time of the Needleman-Wunsch algorithm increases as the length of the sequences in- creases. The time efficiency of the CAT algorithm remains constant regardless of the length of the sequences. Therefore, the advantage of the proposed method is the fast processing in the alignment of large sequences, for which the execution of the exact algorithms takes a long time.

B.4.3. **Gancheva V.**, Jongov T., Georgiev I. Medical X-Ray Image Classification Method Based on Convolutional Neural Networks, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol 13920, pp. 225-244, 2023, **Scopus, SJR = 0.606 (2023), Q3**

Artificial intelligence and machine learning, including convolutional neural networks are increasingly entering the field of healthcare and medicine. The aim of the study is to optimize the learning process of convolutional neural networks through X-ray images pre-processing. A model for optimizing the overall architecture of a classifying convolutional neural network of chest X-rays by reducing the total number of convolutional operations is presented. The method can be applied in any field of image classification. The experimental results of the research prove the successful application of the optimization process on the training of classification convolutional networks, as the optimization does not affect the accuracy of the trained models. There is a significant reduction in the training time of each epoch in the optimized convolutional networks. The optimization is of the order of 25% for the network with an input layer size of 124 x 124 and about 27% for the network with an input layer size of 122 x 122. At the same time, there is no significant deviation in the values of losses and accuracy on training data of the three types of neural networks. The values of losses and accuracy on the validation data show significant variations, which do not give a significant advantage to any of the neural networks, but rather are arbitrary. Three datasets have been used for conducting the research – one dataset for X-ray masks and two datasets for image classification. The images are in lossy jpeg or png format not of perfect quality, but are suitable for proving the point of the research. The current study is applied to segmentation and classification of X-ray images of the lung, but the method can be applied in any field of image classification in which the informative image regions are grouped and subject to segmentation. Of great importance is the stage of preliminary segmentation and analysis of the active regions, in which the distribution of the widths and heights of the obtained active regions is studied. This distribution is necessary to determine the effectiveness of the described optimization model.

B.4.4. **Gancheva V.**, Georgiev I. A Scalable Healthcare Data Science Framework Based on Service-Oriented Architecture, In Proc. of International Conference on Research in Education and Science, May 18-21, 2023, Cappadocia, Turkiye, pp. 2525-2536, **Scopus, SJR=0.106 (2023)**

The aim of the research presented in this paper is to propose a conceptual model and architecture of a service-oriented scalable framework, ensuring the implementation and verification of methods and algorithms for the integration, management, analysis, and visualization of biomedical data and the implementation of scientific research for the needs of precision medicine. The system architecture for big biomedical data analytics and discovering useful knowledge from data consists of the following components: biomedical data sources, data storage, data integration and preprocessing, real-time data flow, stream processing, analytical data storage, data modeling and analysis, and results visualization. Parallel processing and streaming technologies are used. Multiple data can be updated in the research data warehouse for analysis and visualization. A feed-forward artificial neural network is designed for data analysis, and during the training process, the input data is divided into training data and test data. The training error and its distribution over the weights of the neurons in the network are determined. A reduced set of statistical records related to cardiovascular disease analysis has been used as experimental data. The original database contains 76 attributes, and 14 of them have been used for the study. In addition, the data is split in a ratio of 0.8 to 0.2. The first 80% of the data was used to train the neural network and the remaining 20% to test the trained network. The calculated accuracy increases with increasing epochs and is higher for the training data and lower for the validation test data. Thus, the trained model can be saved, and loaded on another system, as well as available for review of the weight values. The trained model is applied in the system to calculate new input parameters that were not used either in training or validation.

B.4.5. **Gancheva V.** Platform for Big Biomedical Data Streams Management and Analytic, *International Journal of Circuits, Systems and Signal Processing*, pp. 580 - 588, ISSN 1998-4464, **Scopus, SJP = 0.156 (2020), Q4**

A platform for multidimensional large-scale biomedical data management and analytics is presented in this paper. The goal is to suggest an intelligent solution as integrated, scalable workflow development environment consisting of a suite of software tools to automate the computational process in conducting scientific experiments.The suggested platform aims intelligent data management, analysis and visualization. The advantages in data management, analysis, knowledge discovery and visualization empower the scientists to achieve new scientific breakthroughs. As a result, the research work is directed towards developing computer aided diagnostic system for solving problems in the field of precision medicine. A breast cancer prediction algorithm based on machine learning is presented. The research techniques follow the processing pipeline for discovering useful knowledge from a collection of data and cover the following: data preprocessing; knowledge discovery and decision making; comprising results and interpreting accurate solutions from the observed results. Four machine learning algorithms for breast cancer classification are selected and evaluated experimentally: Random Forest, kNN, Logistic Regression, and SVM.

B.4.6. **Gancheva V.**, Borovska P. SOA Based System for Big Genomic Data Analytics and Knowledge Discovery, Proceedings of 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), DOI: 10.1109/IDAACS.2019.8924370, **Scopus/ Web of Science**

SOA based system for adaptive knowledge discovery and decision-making based on big genomic data analytics is proposed in this paper. The system architecture is comprised of web services for data integration, preprocessing of large data streams, knowledge discovery based on genomic data analytics, knowledge interpretation and results visualization. The designed system architecture is comprised of web services for: (1) searching and integration of heterogeneous data from different data sources and in various formats; (2) data preparation, cleansing, filtering and selection; (3) data processing and analyses, and (4) knowledge representation and results visualization. The future work is to make experiments thought the system in the area of molecular biology. The spectrum of case studies comprises identifying regulatory elements in sequenced genomes, and prediction of the type and malignance of breast cancer. This will enable fast processing of clinical observations and laboratory analyses data and comparison with the available data accumulated so far in support of precision medicine.

B.4.7. **Gancheva V.**, Georgiev I. Software Architecture for Adaptive In Silico Knowledge Discovery and Decision Making Based on Big Genomic Data Analytics, AIP *Conference Proceedings 2172*, 090009 (2019), International Conference on Application of Mathematics in Engineering and Economics (AMEE'19), AIP Conference Proceedings, **Scopus/ Web of Science, SJR = 0.190 (2019)**

Software architecture for adaptive knowledge discovery based on big genomic data analytics is presented in this paper. The software architecture is comprised of layers for data integration and preprocessing, database/data warehouse server, data discovery engine, pattern evaluation and graphical user interface. The big genomic data architecture consists of data sources, storage, integration and preprocessing, real data stream, stream processing, analytical data store, analysis and reporting. An algorithm for prediction of breast cancer based on machine learning for processing and analysis of big genomic data and knowledge discovery with respect to personalized treatment is presented. The proposed algorithm for breast cancer classification is implemented using Stochastic Dual Coordinate Ascent (SDCA) method and Wisconsin breast cancer database. Experimental results are presented and discussed. The purpose of the study is to apply the software architecture on big genomic data analytics by practical experiments for specific case study identifying regulatory genetic elements in sequenced genomes, and prediction of the type and malignancy of breast cancer. This will enable fast processing of clinical observations data and comparison with the available data accumulated so far in support of precision medicine. The proposed software architecture is based on machine learning and procedures for generating models and rules tailored to the target of scientific research. An integrated approach for support of the knowledge discovery, based on adaptive machine learning and adaptive procedures for generating rules according to the goal of scientific research is explained. The advantage of the proposed framework is automatic generation of the hypothesis and options for decisions making on the basis of the learning data set analysis, while the verification and validation is conducted via benchmark testing data set and the expertise of the researchers of the relevant area.

B.4.8. **Gancheva V.**, Stoev H. DNA Sequence Alignment Method Based on Trilateration, Bioinformatics and Biomedical Engineering, *Lecture Notes in Computer Science*, 2019, vol. 11466, Springer, Cham, pp. 271-283, **Scopus/ Web of Science, SJR = 0.427 (2019), Q2**

The effective comparison of biological data sequences is an important and a challenging task in bioinformatics. The sequence alignment process itself is a way of arranging DNA sequences in order to identify similar areas that may have a consequence of functional, structural or evolutionary relations between them. The goal of this paper is to present a new effective and unified method for sequence alignment on the basic of trilateration method. This method suggests solutions to three major problems in sequence alignment: creating a constant favorite sequence, reducing the number of comparisons with the favorite sequence, and unifying / standardizing the favorite sequence by defining benchmark sequences. An innovative method for searching DNA sequences based on the trilateration method is proposed in this paper. Three constant benchmarks for the trilateration implementation have been defined, which create a constant favorite sequence, i.e. it does not depend on the data in the database and change remains the same. This allows making comparisons at the outset – during input of the sequences in the database and it can be recorded as metadata to each sequence. Thus, there is no need to make a comparison of the sequences during the search, but instead will only compare the metadata. By establishing benchmark sequences have been solved the problem of unification / standardization of sequence favorite for all facilities using the described algorithm to compare. Calculations suggested in proposed algorithm is relatively simple and fast, making it suitable for use as a first step in biological sequences alignment algorithms.

B.4.9. Borovska P., **Gancheva V.**, Georgiev I. Platform for Adaptive Knowledge Discovery and Decision Making Based on Big Genomics Data Analytics, Bioinformatics and Biomedical Engineering, *Lecture Notes in Computer Science*, 2019, vol. 11466. Springer, Cham, pp. 297-308, **Scopus/ Web of Science, SJR = 0.427 (2019), Q2**

In the past years, researchers and analysts worldwide determine big data as a revolution in scientific research and one of the most promising trends that has given impetus to the intensive development of methods and technologies for their investigation and has resulted in the emergence of a new paradigm for scientific research Data-Intensive Scientific Discovery (DISD). The paper presents a platform for adaptive knowledge discovery and decision making tailored to the target of scientific research. The major advantage is the automatic generation of hypotheses and options for decisions, as well as verification and validation utilizing standard data sets and expertise of scientists. The platform is implemented based on scalable framework and scientific portal to access the knowledge base and the software tools, as well as opportunities to share knowledge and technology transfer. In this paper, a platform for adaptive knowledge discovery and decision-making based on big data analytics is proposed. The major advantage is the automatic generation of hypotheses and options for decisions, as verification and validation are performed using standard data sets and expertise of scientists. The tools for utilizing the platform are scalable framework and scientific portal to access the knowledge base and the software tools, as well as opportunities to share knowledge, and technology transfer. Web portal provides services to access and extract knowledge out of biological data and execute parallel software applications for big genomics data analysis. An integrated approach to support knowledge discovery and decision-making based on big data analytics, adaptive machine learning and adaptive procedures for generating rules according to the goal of scientific research is presented. The future work is to make in silico experiments on the platform based on big genomic data

analytics for scientific research in the area of molecular biology. The spectrum of case studies under investigation comprises identifying regulatory elements in sequenced genomes, and prediction of the type and malignance of breast cancer. This will enable fast processing of clinical observations and laboratory analyzes data and comparison with the available data accumulated so far in support of precision medicine.

B.4.10. Borovska P., **Gancheva V.**, Georgiev I., Ivanova D., Hybrid Parallel Multiple Sequence Alignment Based on Artificial Bee Colony on the Supercomputer JUQUEEN, Proceedings of International Conference on Electrical Engineering and Computer Science (EECS), Bern, Switzerland, 2017, pp. 47-51, ISBN: 978-1-5386-2086-1, DOI: 10.1109/EECS.2017.18, **Scopus/ Web of Science**

The paper focuses on performance investigation and improvement of multiple biological sequence alignment software MSA_BG on the BlueGene/Q supercomputer JUQUEEN. For this purpose, scientific experiments in the area of bioinformatics have been carried out, using as case study influenza virus sequences. The parallel software MSA_BG for multiple sequence alignment has been ported and tuned on the Blue Gene/Q supercomputer JUQUEEN. The objectives of the investigation are code optimization, porting, scaling, profiling and performance evaluation of MSA_BG software. To this end we have developed hybrid MPI/OpenMP parallelization on the top of the MPI only code and we showcase the advantages of this approach through the results of benchmark tests, performed on JUQUEEN. The experimental results show that the hybrid parallel implementation provides considerably better performance than the original code. Hybrid MPI/OpenMP parallelization was implemented and evaluated experimentally. Parallel performance was investigated and optimized through benchmark tests and profiling. The implementation of hybrid MPI/OpenMP parallelization on MSA_BG reduces up to a good factor the overall runtime of the MPI only version of the application as it allows us to fully occupy the CPU capacity of a JUQUEEN node with threads. It should also be noted that runs using the hybrid implementation resulted in better quality of sequence alignments due to additional randomness that was produced by the Mersenne Twister generator. The performance estimation and analyses show that the hybrid parallel program implementation of MSA_BG algorithm scales well as the number of the cores increases and is well balanced both in respect to the workload and machine size. The optimized code is universal and can be applied for other similar research projects and experiments in the field of bioinformatics. MSA_BG software allows researchers to conduct their experiments and perform simulations with very large amounts of data.

B.4.11. Borovska P., **Gancheva V.**, Landzhev N. Massively Parallel Algorithm for Multiple Biological Sequences Alignment, Proceedings of the IEEE International Conference on Telecommunications and Signal Processing (TSP), Rome, Italy, ISBN 978-1-4799-0402-0, pp. 638-642, DOI: 10.1109/TSP.2013.6614014, **Scopus/ Web of Science**

In silico biological sequence processing is a key for molecular biology. This scientific area requires powerful computing resources for exploring large sets of biological data. Multiple sequence alignment is widely used method for biological sequence processing. The goal of this method is DNA and protein sequences alignment. This paper presents an innovative parallel algorithm MSA_BG for multiple alignment of biological sequences that is highly scalable and locality aware. The designed MSA_BG algorithm is iterative and is based on the concept of Artificial Bee Colony metaheuristics and the concept of algorithmic and architectural spaces correlation. The metaphor of the ABC metaheuristics has been constructed and the functionalities of the agents have been defined. The conceptual parallel model of computation has been designed. The algorithmic

framework of the designed parallel algorithm has been constructed. The choice of the algorithm ABC is based on the fact that in essence it is a hybrid metaheuristics - a combination of methods based on populations (scouts generate simultaneously a number of possible solutions) and a method based on trajectories (employed bees perform the local searches around the decisions of the scouts, seeking to improve the decisions quality). MSA_BG algorithm has a hierarchical structure, which enables observing the principle of locality (independent calculations), and very high scalability (hives and swarms), so it is expected high efficiency implementations for petaflops supercomputers. Parallel performance evaluation and profiling of multiple sequence alignment on the basis of MSA_BG algorithm on supercomputer BlueGene/P have been proposed in this paper. The case study is investigating the viral nucleotide sequences and finding out consensus motifs and variable domains in the different segments of the influenza virus. Parallel performance parameters such execution time, accelerating time and profiling have been estimated experimentally. The performance estimation and profiling analyses have shown that the parallel system is well balanced both in respect to the workload and machine size except for the process rank 0 that is the most heavily utilized due to performance of data distribution to all other processors, communication and synchronization.

Publications outside the habilitation work

Г.7.1. Trendafilov I., Gancheva V. Neuromorphic Assisted Sensor Grids, XXXII International Scientific Conference Electronics, 2023, DOI: 10.1109/ET59121.2023.10279437, Scopus

A particular problem in using artificial intelligence techniques in the sensor grid is the high power consumption. Remote sensors are usually limited by the amount of power available, thus our general goal is to minimize it while preserving accurate experimentation results. Using emerging technologies like spiking neural networks and neuromorphic hardware, we can create sensor grids that perform data processing with preserving low power consumption. In this paper, we demonstrate a remote node with integrated visual sensor that works on less than 10mW of energy, while performing continuous scene monitoring, object detection and classification. The system has intelligent power management and the ability to send data wirelessly.

Г.7.2. Trendafilov I., Gancheva V. Neuromorphic Neurons and Networks for Artificial Intelligence Built Using Temporal Space Calculations, XXXII International Scientific Conference Electronics, 2023, DOI: 10.1109/ET59121.2023.10279371, Scopus

We deigned a spiking neural network that computes network weights in the temporal dimension. Such a network can be used for artificial intelligence and deep learning. We demonstrate circuits implementing blocks for building such a network and then a training model. This enables creation of efficient Hessenstein-Reichardt detectors observed in motion detection networks in the nature. The proposed network allows reconfiguration across network layers algorithmically. We had designed a novel spiking neural network that simplifies the system parameters and reduces the signal dimensionality to binary signaling. This type of networks should be stable and easy to implement in neuromorphic hardware enabling the building of very large networks with billions of neurons. The training algorithm allows for network reconfiguration, a behavior observed in the nature. Further investigations is required but our working hypnotizes is that such a network can be used to implement any task that requires previous state of any neuron in the network. In this paper we had demonstrated the technical viability of the method,

therefore we can proceed with software based simulations of such networks to prove this. We demonstrated that we can use the charge as an intermediate quantity variable. We had used capacitors for storing the electric charge, but in future work we plan to investigate the usage of memristors in this role.

Г.7.3. Trendafilov I., **Gancheva V.** Hassenstein-Reichardt Detector Using Controllable Single Pulse Time-Delay Circuit for Neuromorphic Hardware, International Scientific Conference Computer Science, 2023, DOI: 10.1109/COMSCI59259.2023.10315865, **Scopus**

Inspired by the visual system of the fruit fly, a common building block for neuromorphic hardware is created, which is vital for third-generation neural networks by allowing the delay to be parameterized in an efficient way. This allows the delay to be parameterized in an inexpensive way. The advantage of our proposal is the ability to build and test new networks using a dynamic approach in artificial intelligence. A circuit is designed that creates a pulse delay line with controllable time parameters that can be used to build Hassenstein-Reichardt detectors and integrate into neuromorphic hardware operating with pulsed neural networks. The parameters can be changed during training of the neural network. The circuit is implemented by using two capacitors, each paired with a controllable voltage source. This provides two independent timing parameters. The first capacitor has a charge proportional to the length of the input pulse, while the second sets the delay between the input and the output.

Г.7.4. **Gancheva V.** Application of Machine Learning Techniques for Software Anomaly Detection, International Conference on Applied Mathematics & Computer Science (ICAMCS), Lefkada Island, Greece, August 8-10, 2023, IEEE Catalog Number: CFP23T98-ART, ISBN: 979-8-3503-2426-6, DOI: 10.1109/ICAMCS59110.2023.00016, **Scopus**

A rising variety of platforms and software programs have leveraged repository-stored datasets and remote access in recent years. As a result, datasets are more vulnerable to malicious attacks. As a result, network security has grown in importance as a research topic. The usage of intrusion detection systems is a well-known strategy for safeguarding computer networks. The research presented in this paper proposes a hybrid anomaly detection method that blends rule-based and machine-learning-based methods. The advantage of the proposed methodology is the combination of different methods and algorithms. In order to construct the appropriate rules, a genetic algorithm is utilized. Principal component analysis is used to extract the relevant features aimed to improve the performance. The suggested method is validated experimentally using the KDD Cup 1999 dataset, which meets the requirement of using appropriate data. The proposed software anomaly detection method is verified experimentally by implementing three classification algorithms. An analysis and evaluation of the obtained results in terms of accuracy and precision were made. The suggested solution is used to identify and examine four different types of assaults in a well-known benchmark dataset: Neptune, Ipsweep, Pod, and Teardrop. The KDD Cup 1999 dataset, which satisfies the condition of using acceptable data, is used to empirically validate the suggested method. The KDD Cup 1999 dataset consists of 41 features that are broken down into required, traffic, and content aspects, as well as training and test data. The KDD Cup 1999 dataset contains roughly five million raw data points, with attack data accounting for about 80% of these. After testing the characteristics specified in the training phase, the data is classified into attack categories and normal behavior during the machine learning phase.Four "attack" groups and one "normal" category comprise these statistics. Experiments are performed based on Support Vector Machine, Decision Tree, and Naive Bayes algorithms and are aimed at accuracy and probability in the analysis of datasets. The analysis done

shows the best results in the case of the Naive Bayes classification algorithm and can be assumed to be the most reliable in comparison with the results in the cases of Support Vector Machine and Decision Tree.

Г.7.5. Draganov I., Gancheva V. Optimizing the Non-local Means Filtering of CT Images. Medical Imaging and Computer-Aided Diagnosis. MICAD 2022. *Lecture Notes in Electrical Engineering*, vol 810. Springer, Singapore, https://doi.org/10.1007/978-981-16-6775-6_1, Scopus, SJR=0.147 (2022), Q4

In this paper a general optimizing procedure is proposed for the non-local means (NLM) filter. It involves finding the optimal degree of smoothing, the size of the search window and the size of the comparison window for a series of Computed Tomography (CT) images. All of them contain Additive White Gaussian Noise (AWGN) with a particular variance and zero mean, both of which are preliminary unknown. Applying the optimization procedure over a single slice from the CT packet appears to be efficient enough in finding the optimal parameters of the filter for the rest of the CT images. Positive results are obtained from filtering a complete set of CT images from a patient's body and the quality of the filtration is higher than that of the Gaussian and Average filters. Experimental results show that the Degree of Smoothing (DoS) affects the quality of the reconstructed images. The increase of this parameter leads to saturation of both the Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). There is a minimal value for DoS which could be found as an optimal at the beginning of the saturation zone. The change of DoS has no significant effect on the filtration time. The sizes of the search and comparison windows also have non-linear effect over the quality of the reconstructed images. For both of them there are saturation areas in the PSNR and SSIM functions. It is possible to select the minimum windows sizes, such that they lay at the beginning of the saturation zone. Thus, they guarantee best quality of the images at the lowest computational time. The computational time, itself, increases monotonically with the increase of the surface of the search and comparison window. The NLM filter provides better quality of the filtered CT images than the Gaussian and Average filters for wide range of noise level of AWGN. The filtering time of all three filters does not depend on the noise level. The NLM filter is more than 2 orders of a magnitude slower than the other two filters. There is no grainy structure in the images, filtered by the NLM, but there is a little loss of contrast. As a future work optimization of the NLM filter as processing time could be undertaken.

Г.7.6. Ganheva V., Todorova V. Workflow for Medical Data Classification and Analysis, 6th International Symposium on Multidisciplinary Studies and Innovative Technologies, October 20-22, 2022, Ankara, Turkey, DOI: 10.1109/ISMSIT56059.2022.9932780, Scopus

An approach for automated knowledge extraction and decision-making from medical images through a workflow for preprocessing of incoming X-ray images, analysis, classification and evaluation of the results is presented in this paper. The designed algorithm for analysis of medical X-rays images is based on machine learning and consists of three main phases: preprocessing of training and validation datasets, medical images classification utilizing Logistic Regression, Naïve Bayes, SVM methods, evaluation of the model. A workflow was developed to process and analyze datasets of lung X-ray images containing four classes, and determine classification accuracy by examining performance evaluation parameters. The analysis performed shows the advantage of the Logistic Regression results, which are assumed as better comparing with results obtained by Naive Bayes and SVM. The attribute Category is selected as the target for the classification. Samples 66% of the data were selected as the training dataset. The remaining data is used as the test dataset.

Г.7.7. Ganheva V., Vetova S. Approach and Concept of Workflow for Animal Husbandry Data Integration and Analysis, 30th National Conference with International Participation (TELECOM), 27 - 28 October 2022, Sofia, Bulgaria, pp. 1-4, DOI: 10.1109/TELECOM56127.2022.10017334, Scopus/ Web of Science

A concept for data integration and risk analysis in animal husbandry production is presented in this paper. The proposed animal husbandry data integration and analysis model structure consists of three layers, each of which combines the tasks to be performed. The architecture for data integration and its components are described. The workflow organization is depicted in details including modules and their connections used for data exchange. On the base of the presented workflow, experiments were implemented. To perform the experimental part of the study, statistical data on the population of domestic animals is used. The analysis of their results show the trend in the risk of animal husbandry production. The model demonstrates the ability to load process and analyze the 141 records initially defined in the original statistics CSV file. The statistics provide information for analyzing the risk of extinction of the selected species of domestic animals by breed. The created model for processing and analyzing statistical data for assessing the risk of extinction of animal breeds is applicable in the field of animal husbandry. It allows tracking the growth trend of a given breed over a period of time based on pre-collected statistical data. Based on the automated processing of the data and the assessment of the risk of extinction and conclusions, as a next step measures can be taken to protect animal breeds. The statistics provide information for analyzing the risk of extinction of the selected species of domestic animals by breed.

Г.7.8. Draganov I., GanchevaV. Unsharp Masking with Local Adaptive Contrast Enhancement of Medical Images, *Lecture Notes in Electrical Engineering*, vol 784. Springer, Scopus, SJR = 0.147 (2022), Q4

In this paper we present a generalized algorithm for unsharp masking of medical images, which takes as one of its inputs a high contrast image, underwent local adaptive contrast enhancement. Selection of optimal values of the number of histogram bins, processing window size and intensity lower and upper limits in iterative manner is part of applying Contrast Limited Adaptive Histogram Equalization (CLAHE). Optimization procedures for histogram equalization, intensity adjustment, and contrast constrained histogram equalization algorithms are presented to find optimal parameters for them. Root mean square contrast, sharpness, and structural similarity between a contrast-enhanced image and the original image play the role of target parameters. Experimental results reveal higher quality of the output images in terms of both root mean square contrast and sharpness. Achieved quality, both visually and quantitatively, is compared to that from the Adaptive Histogram Equalization (AHE) algorithm, limited histogram stretching and ordinary histogram equalization, which proves its applicability. Tests with CT and X-ray images confirm the plausibility of the approach taken and the applicability of the resulting images for the roughness-masking algorithm to use as input. Contrast-limited adaptive histogram equalization yields more detailed and contrast-enhanced final images followed by histogram equalization and image adjustment algorithms at the cost of more computational time. Unsharp masking in this general and easy-to-perform form is considered a useful tool for medical purposes. The algorithm is considered appropriate for processing a number of types of images, such as CT, X-ray, etc.

Г.7.9. Draganov I., **Gancheva V.** Optimal Bilateral Filtering of CT Images, International Conference on Computational Science and Computational Intelligence (CSCI), 2021, pp. 1668-1672, DOI: 10.1109/CSCI54926.2021.00053, **Scopus/ Web of Science**

In this paper a general optimization algorithm is proposed for tuning the parameters of a bilateral filter when processing Computed Tomography images, containing Additive White Gaussian Noise. The Peak Signal to Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM) are the target parameters during the optimization with the explicit aim of achieving their maximums. The operation of the optimal configuration of the bilateral filter is compared with the results from filtering of the same images with the Gaussian and average filters. Positive results are obtained and the proposed optimization is considered applicable not only for Computed Tomography images, but also for Magnetic Resonance Imaging, multispectral and for hyperspectral images. The objective quality parameters depend, both on the coverage and on the intensity range used, the dependence of which becomes stronger as the variance of the available noise increases. Applying the optimization procedure to a slice of a CT image and subsequently filtering all slices provides an efficient way to obtain the highest quality for the entire set. Future work on the topic will reveal its applicability not only to different types of images, but also to further refine the optimization for different types of noise using suitable adapted filter shapes.

Г.7.10. **Gancheva V.** Parallel Multithreaded Medical Images Filtering, International Conference on Computational Science and Computational Intelligence (CSCI), 2021, pp. 1788-1793, DOI: 10.1109/CSCI54926.2021.00338, **Scopus/ Web of Science**

The quality of medical images is paramount. Being of high grade, it guarantees the quality of medical diagnosis, treatment and quality of patient's life through the means of health care or using automate intelligent systems for medical diagnosing, treatment and monitoring. The paper presents the computational challenges in medical images processing. The great challenges are to propose parallel computational models and parallel program implementations based on the algorithms for medical images filtering. Parallel computational model based on two-dimensional filters is designed. The proposed parallel model is verified by multithreaded parallel program implementation. An investigation of the efficiency of medical images filters based on parallel multithreaded program implementation, applying twodimensional filters on a given list of compressed jpeg medical images and generating output jpeg images for each type of applied filter. The applied filters are Brightness Control, horizontal and vertical filter of Sobel, Laplace and Blur. A number of experiments have been carried out for the case of dataset consisted of 162 whole mount slide images of Breast Cancer (BCa) specimens scanned at 40x and various number of threads. Parallel performance parameters execution time and speedup are estimated experimentally. The performance estimation and scalability analyses show that the suggested model has good scalability.

Г.7.11. Ko S.-H., **Gancheva V.** An Approach for Parallel Reading in Multiple Sequence Alignment, International Conference Automatics and Informatics, ICAI 2020, DOI: 10.1109/ICAI50593.2020.9311347, **Scopus**

We propose an approach for faster file reading of multiple sequence alignment input through the use of MPI-I/O over a subset of MPI cores. The idea is to let a subset of MPI cores to perform the I/O operation and locally broadcast to individual neighbors so that the code is less sensitive to the stability of the parallel file system. It is achieved by creating a number of subgroups under a global MPI communicator. The size of each subgroup and the buffer size of each reading operation are tuned through the synthetic benchmark. We verify the performance of our approach by comparing it with the traditional way of

“sequential file reading and global broadcast”, and apply it to the MPI version of multiple sequence alignment software ClustalW. We divide total number of MPI rank to sub-groups and let the local master of each group to conduct I/O operations, instead of opening the I/O protocol to all ranks. Along with this group size, the acquired data size per a single read instruction also affects much on I/O performance. We performed extensive benchmark experiments to determine the optimal values of these two parameters. Benchmark runs result that the best performance is achieved when the group size is 1/4th or 1/8th of total number of processors and the reading chunk size is set as the file size. In that configuration, parallel I/O outperforms the serial I/O by 2 – 4 times at tens or hundreds of MegaByte datasets. The application to the ClustalW-MPI soft-ware results that the time for storing the sequence data is accelerated by 6.8 times by the adoption of parallel I/O with 8192 BlueGene/Q cores. We argue that the current parallel I/O interface design can provide much gain in I/O performance of many bioinformatics softwares.

Г.7.12. Aleksieva-Petrova A., Gancheva V., Petrov M. Software Architecture for Adaptation and Recommendation of Course Content and Activities Based on Learning Analytics, Proceedings of International Conference on Mathematics and Computers in Science and Engineering, DOI: 10.1109/MACISE49704.2020.00010, Scopus/ Web of Science

Nowadays the main challenge in learning analytics is to suggest efficient methods and technologies in order to achieve better learner results. This paper presents a software architecture for adaptation and recommendation of course content and activities based on learning analytics. It is comprised of layers for ingestion layer, aggregation layer, storage layer and big data processing and analyses layer. An algorithm for prediction of student learning based on machine learning for processing and analysis of data and knowledge discovery with respect to main learner and teacher activities is presented. The proposed algorithm for student learning classification is implemented by using Averaged Perceptron method. Experimental results are presented and discussed. The purpose of the study is to apply the software architecture on learning analytics by practical experiments for specific case study identifying event elements in sequenced learners' and courses' activities logs, and student learning prediction. The real time learning and gaming analytics of big data produced by modern e-learning platforms and educational games, for a learner-centric adaptation of technology enhanced learning is one of main challenge. The system helps to structure and storage of big data from heterogeneous sources as both LMS and educational game; identify patterns by analysing learners’ behaviour and allowing data analyses with descriptive, predictive, and prescriptive results. The experimental data set is obtained from learning management system and contains of 63774 instances characterized by 7 attributes.

Г.7.13. Gancheva V. Knowledge Discovery Based on Data Analytics and Visualization Supporting Precision Medicine, International Conference on Mathematics and Computers in Science and Engineering, pp. 102 - 105, DOI: 10.1109/MACISE49704.2020.00024, Scopus/ Web of Science

A comprehensive system for precision medicine, which covers all phases of data discovery, data integration, data preprocessing, building models, data storage, data analysis and visualization can be very useful to scientists in support of precision medicine. The software system aims intelligent big genomic data management, analysis and visualization and allows scientists an easy, fast and flexible approach for data processing. They can choose the services they wish to be executed, use the available data sets in databases, or enter their own data to be processed. A software application for biological data visualization has been developed for the purpose of system testing and validation. The proposed application provides an opportunity for three-dimensional visualization of the proteins structure or DNA sequence implemented through OpenGL. The three-dimensional modeling of the corresponding macromolecules

enables one to gain a clear view of the objects complexity at the atomic level. Complex molecules can be displayed by using modern technologies for 3D modeling.

Г.7.14. Gancheva V., Georgiev I. Multithreaded Parallel Sequence Alignment Based on Needleman-Wunsch Algorithm, Proceedings of 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), DOI: 10.1109/BIBE.2019.00037, Scopus/ Web of Science

Biocomputing and molecular biology are areas that change knowledge and skills for acquisition, storing, management, analysis, interpretation and dissemination of biological information. This requires the utilization of high performance computers and innovative software tools for management of the vast information, as well as deployment of innovative algorithmic techniques for analysis, interpretation and prognostication of data in order to get to insight of the design and validation of life-science experiments. Sequence alignment is an important method in DNA and protein analysis. The paper describes the computational challenges in biological sequence processing. The great challenges are to propose parallel computational models and parallel program implementations based on the algorithms for biological sequence alignment. An investigation of the efficiency of sequence alignment based on parallel multithreaded program implementation of Needleman-Wunsch algorithm is presented in this paper. Parallel computational model based on Needleman-Wunsch algorithm is designed. The proposed parallel model is verified by multithreaded parallel program implementation utilizing OpenMP on an 8 cores server Xeon. A number of experiments have been carried out for the case of various data sets and a various number of threads. Parallel performance parameters execution time and speedup are estimated experimentally. The performance estimation and scalability analyses show that the suggested model has good scalability both in respect to the workload and machine size, and scales better as the number of the cores increases.

Г.7.15. Gancheva V. A Big Data Management Approach for Computer Aided Breast Cancer Diagnostic System Supporting Precision Medicine, *AIP Conference Proceedings 2172*, 090012(2019), International Conference on Application of Mathematics in Engineering and Economics (AMEE'19), AIP Conference Proceedings, Scopus/ Web of Science, SJR = 0.190 (2019)

An approach to management of large amount of heterogeneous data sets from various data sources for a breast cancer diagnostic system is presented in this paper. Big genomic data architecture consists of data sources, storage, integration and preprocessing, real data stream, stream processing, analytical data store, analysis and reporting. Activities at data management for breast cancer diagnostic system are explained. Conceptual database architecture for storing data sets of several types in order to support breast cancer prediction is designed. The breast cancer database comprises of information related to breast cancer genes and functions - id, name, type, organism, function, and proteins coded, description, link for retrieving sequence. The patient's database consists of individual patient data - genetic data, clinical history, individual life style parameters, clinical tests results, environmental factors. The data sets in the suggested big data management system are retrieved from the biomedical research databases. The data management system is platform independent, easy to use and provides access to other databases such PubMed, NCBI. The purpose is to be used for data storage in a system for big data analytics and knowledge discovery, especially for the case study of breast cancer diagnostic. The advantages in data management, analysis, and knowledge discovery empower the scientists to achieve new scientific breakthroughs. As a result the research work is directed towards rapid management and processing of clinical data for solving problems in the field of precision medicine.

Г.7.16. **Gancheva V.** SOA Based Multi-Agent Approach for Biological Data Searching and Integration, *International Journal of Biology and Biomedical Engineering*, ISSN: 1998-4510, Volume 13, 2019, pp. 32-37, Scopus, SJR = 0.191 (2019), Q4

Models for extraction and integration of large amount of genomics data, as well as problems related to heterogeneity, distribution and compatibility of data are presented in this paper. SOA based multi-agent approach for biological data searching and integration is proposed. A conceptual architecture for integrating of distributed biological data based on SOA is designed. The architecture is aimed to automate the data integration and allows the rapid management of large volumes of diverse data sets represented in different formats - relational, NoSQL, flat files. The integration of different databases is solved by using multi-agent architecture. The integration system consists of services for transforming the common request into a specific language request for each local database, depending on its type. The conceptual database integration is solved by applying translating query approach. Each integrated database is represented by a separate conceptual scheme called a virtual scheme. This scheme is generated in the collating process, which compares structural elements from the database to the conceptual model. Service oriented multi-agent system for searching of biological data from different sources that sends queries to multiple databases and then compiles the results into a list, depending on the type of source is developed. The system allows the user to set search criteria and access multiple databases simultaneously. The services allow the system to be accessed over the Internet by multiple clients (mobile phones, web browsers, desktop applications) and serving a wide range of users simultaneously.

Г.7.17. Borovska P., **Gancheva V.**, Ko S.-H., Scaling of Parallel Multiple Sequence Alignment on the Supercomputer JUQUEEN, Proceedings of the International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications IDAACS'2013, Berlin, German, ISBN 978-1-4799-1426-5, pp. 687-691, DOI: 10.1109/IDAACS.2013.6663013, Scopus

In this paper is proposed optimization, scaling, performance evaluation and profiling of parallel multiple sequence alignment based on ClustalW algorithm on the supercomputer BlueGene/Q, so-called JUQUEEN, for the case study of the influenza virus sequences. For this purpose, a parallel I/O interface for simultaneous and independent access to single file collectively has been designed and verified based on parallel program implementation on the supercomputer JUQUEEN. Parallel performance evaluation and profiling of multiple alignment based on parallel program implementation of ClustalW algorithm utilizing Message passing interface on the supercomputer JUQUEEN have been proposed in this paper. Parallel performance parameters such execution time, scaling and profiling have been estimated experimentally. The performance estimation and profiling analyses have shown that the parallel system is well balanced both in respect to the workload and machine size except for the process rank 0 that is the most heavily utilized due to performance of communication and synchronization. The parallel I/O implementation for acquiring the sequence dataset results in the strong speed-up for initial procedure. It reaches up to 6.8 times faster than the baseline code in case of 8K cores on JUQUEEN. Yet, the overall gain by the parallel read is not strong since the input file size is limited due to allocated memory size. JUQUEEN's memory allocation per core is 1GB, which limits maximum number of input sequences to approximately 10000 (depending on the sequences length).

Г.8.1. **Gancheva V.** Intelligent Management and Analytics of Big Biomedical Scientific Research, Proceedings of XV International Scientific Conference e-Governance and e-Communications, Sozopol, June 2023

The accumulation and storage of huge amounts of data becomes a major source of knowledge. The new paradigm for research implies a new way of conducting experiments and discovering knowledge. In doing so, the data are analyzed and hidden models, significant correlations and cause and effect relationships are sought using intelligent methods for discovering new knowledge. The paper presents the challenges in creating an integrated open technology platform for implementing intelligent solutions for management and analysis of multidimensional big biomedical data, automating efficient methods and algorithms for the analysis of big biomedical data and applying models for their visualization. The main objective of the platform is to support knowledge extraction and decision-making for the needs of medicine and biology by offering an integrated solution for management, storage, analysis and visualization of large sets of heterogeneous data and providing an easy-to-use infrastructure for conducting scientific research and increasing efficiency.

Г.8.2. Todorova V., **Gancheva V.**, Mladenov V. COVID-19 Medical Data Integration Approach, *Journal Molecular Sciences and Applications*, Volume 2, pp. 102-106, ISSN / EISSN : 2944-9138 / 2732-9992, DOI: 10.37394/232023.2022.2.11

A conceptual model for medical data integration and processing consisting of three layers and six phases is proposed. Data management consists of three main phases: data preparation for analysis, interpretation and visualization, and the preparation phase includes collection, storage, integration of medical data. The second layer of data analysis involves the application of medical data processing methods. The data processing process involves manipulating the collected data and performing functions and operations to extract meaningful information such as validation, sorting, summarizing, analysis, reporting, classification. The medical data classification phase involves the process of arranging data into groups based on predetermined criteria. Clustering methods and techniques such as k-Nearest Neighbor (kNN), kMeans, Support Vector machine (SVM), Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Naive Bayes, etc. are applied for data purposes. A medical data integration workflow was designed, including steps to integrate, filter, aggregate, and sort data. The proposed workflow was validated for SARS-CoV-2 medical data from clinical records of 20400 potential patients.

Г.8.3. Borovska P., **Gancheva V.** Massively Parallel Multiple Sequence Alignment on the Supercomputer JUQUEEN, *NAUN International Journal of Computers*, Vol. 12, 2018, pp. 1-8, ISSN: 1998-4308

In silico biological sequence processing is a key task in molecular biology. This scientific area requires powerful computing resources for exploring large sets of biological data. Parallel in silico simulations based on methods and algorithms for analysis of biological data using high-performance distributed computing is essential for accelerating the research and reducing the investment. Multiple sequence alignment is a widely used method for biological sequence processing. The paper focuses on performance investigation and improvement of multiple biological sequence alignment software MSA_BG on the BlueGene/Q supercomputer JUQUEEN. Experimental simulations on the basis of parallel implementation of MSA_BG algorithm for multiple sequences alignment have been carried out for the case study of the influenza virus variability investigation. The objectives of the investigation are code optimization, porting, scaling, profiling and performance evaluation of MSA_BG software. A hybrid MPI/OpenMP parallelization has been developed and the advantages of this approach through the results of benchmark tests, performed on JUQUEEN have been shown. The experimental results show that the hybrid parallel implementation provides considerably better performance than the MPI only implementation.

Г.8.4. Ivanova D., Borovska P., **Gancheva V.** Experimental Investigation of Enhancer-Promoter Interactions out of Genomic Big Data based on Machine Learning, *International Journal of Computers*, Volume 3, 2018, ISSN: 2367-8895, pp. 58-62

The main purpose of the paper is to presents the experimental investigation for detection of enhancer-promoter interactions from genomic big data based on machine learning propose a pipeline for detection of enhancerpromoter interactions. It is realized by using Decision Tree and Support Vector Machine classifiers. The experimental framework is based on Apache Spark environment that allows streaming and real time analysis of big data. Machine learning library of Apache Spark (MLlib) is implemented in python programming language for processing genomic big data. To perform the results, the enhancer-promoter interactions GM12878 and K562 datasets are used. Finally, the experimental results are presented and discussed.

Г.8.5. Borovska P., **Gancheva V.** Parallelization and Optimization of Multiple Biological Sequence Alignment Software Based on Social Behavior Model, *International Journal of Computers*, pp. 69-74, ISSN: 2367-8895, Volume 3, 2018,

The huge amount of biological sequences accumulated in the world nucleotide and protein databases leads to the necessity of efficient tools for structural genomic and functional analysis. This scientific area requires powerful computing resources for exploring large sets of biological data. Multiple sequence alignment is an important method in the DNA and protein analysis, and is generally the alignment of three or more biological sequences of similar length. As a result of the processing, homology can be derived and the evolutionary relationships between the sequences can be explored. The goal of this paper is to propose parallelization and optimization of the multiple sequence alignment software MSA_BG in order to improve the performance, for the case study of the influenza virus sequences. The objective is code optimization, porting, scaling and performance evaluation of the parallel multiple sequence alignment software MSA_BG for Intel Xeon Phi (the MIC architecture). For this purpose a parallel multithreaded optimization including OpenMP has been implemented and verified. The experimental results show that the hybrid parallel implementation utilizing MPI and OpenMP provides considerably better performance than the original code.

Г.8.6. Borovska P., **Gancheva V.**, Georgiev I. Hybrid Parallel Implementation of Multiple Sequence Alignment Software ClustalW on Intel Xeon Phi, Proceeding of Sixth International Conference on Advances in Computing, Electronics and Communication - ACEC 2017, Rome, Italy, Page(s) : 47 – 51, Electronic ISBN : 978-1-63248-138-2, DOI: 10.15224/ 978-1-63248-138-2-10

This work is aimed to investigate and to improve the performance of multiple sequence alignment software ClustalW on the test platform EURORA at CINECA, for the case study of the influenza virus sequences. The objective is code optimization, porting, scaling and performance evaluation of parallel multiple sequence alignment software ClustalW for Intel Xeon Phi (the MIC architecture). For this purpose a parallel multithreaded optimization including OpenMP has been implemented and verified. The experimental results show that the hybrid parallel implementation utilizing MPI and OpenMP provides considerably better performance than the original code.

Г.8.7. Ташев Т., Лазарова М., **Ганчева В.**, Иванова В. Обучение по управление на ресурси в предприятията, Национална конференция с международно участие "ОБРАЗОВАТЕЛНИ ТЕХНОЛОГИИ 2014", Каварна, 12-14 септември 2014

Enterprises all over the world implement a system for planning and managing resources (Enterprise Resource Planning - ERP). This necessitates the training of master's degree students in the field at many universities in Europe and the USA. A study and detailed analysis of similar specialties for bachelor's and master's studies in foreign universities was made, with the aim of clarifying the good practices in the education of students. After a thorough analysis and determination of the necessary competencies, knowledge and skills, as well as the need of the business for specialists in the field, a curriculum was created at the Faculty of English Engineering Education. Innovative disciplines have been developed as a result of the training, which provide resource management specialists in enterprises to meet the needs of the labor market in Bulgaria and the EU.

Г.8.8. Borovska P., Dokomes H., **Gancheva V.**, Tsvetanov S. GRID Resource Broker Architecture Based on Metadata Scheduling Model, *Journal Computer & Communications Engineering*, Vol. 7 No 1/2013, pp. 5-12, ISSN 1314-2291

In this paper we suggest GRID resource broker architecture, based on metadata scheduling model. A scenario for scheduling has been built up based on scheduling model comprising 5 fundamental modules – resources, application, performance, performance metrics, scheduling policy and programming model. The metadata sets of the input and the output of the scheduling model have been defined. The suggested architectural design of the GRID resource broker includes 3 basic modules: a mapper, evaluator price/resources and dispatcher. A resource management scenario for GRID has been constructed by defining the interactions of the basic modules of the resource broker with GRID services.

Г.8.9. Borovska P., **Gancheva V.**, Aleksieva-Petrova A., Dokomes H. Virtual Center for In-Silico Science and Technology Transfer ViSTa, *Journal Computer & Communications Engineering*, Vol. 7 No 2/2013, pp. 5-11, ISSN 1314-2291

The purpose of the virtual center for in-silico science and technology transfer is the establishment of a virtual platform for e-Science, which provides electronic resources of science, expertise, computer models and software tools for biosimulations and biological database for molecular biology, genomics, virtual screening for drug design and related fields of life sciences. The infrastructure ensures rich functionality of tools and services, allows for dynamic linking of computational resources, implementation of distributed computations and achieving high performance. The virtual platform enables to build a flexible network of independent units linked by the information technologies to share skills, in-silico science and technology and providing access to non-traditional expertise of other units. The virtual platform ensures user-friendly interface that facilitates the utmost use by biologists, chemists, geneticists and others. Innovative tools such as scientific portals, agent-based technologists and ontologies are used. A working framework for the integration of GRID and CLOUD services is established, such that provide to applications uniform access to data and resources.

Г.8.10. Боровска П., **Ганчева В.**, Цветанов С., Оптимизация и изследване на паралелната производителност на софтуерен пакет GADGET на суперкомпютър BlueGene/P и паралелни системи с GPGPU ускорители, сп. *Автоматика и информатика*, год. XLVII, 2/2013, стр. 19-27, ISSN 0861-7562

The paper presents an optimization approach for the Gadget software package. Based on a static and dynamic analysis of the Gadget program, the weak points were identified and the fragments were determined, which were identified and accordingly implemented on specialized hybrid architectures using accelerators in order to increase performance, and an example of a specific implementation was shown and illustrated. An evaluation of the parallel performance - execution times, scaling, profiling - was made, based on experimental simulations on the BlueGene/P supercomputer and GPGPU parallel systems. Experimental results show that after the complete optimization of the program code, the acceleration increases by approximately 50% with good scalability.

Г.8.11. Borovska P., **Gancheva V.**, Landzhev N. High Performance Grid Environment for Parallel Multiple Biological Sequence Alignment, Proceedings of The Eighth International Multi-Conference on Computing in the Global Information Technology ICCGI'2013, Nice, France, ISBN 978-1-61208-283-7, pp. 82-87

We presented an environment enabling secure access to the grid based services as follows: security, parallel program implementation execution and database access on distributed heterogeneous high-performance grid infrastructure. Web portal provides as services access and extraction of biological data and execution of parallel program implementations based on algorithm for comparative analysis of biological data. The proposed portal is verified experimentally for the case study of investigation the influenza virus variability. An innovative parallel algorithm MSA_BG for multiple alignment of biological sequences that is highly scalable and locality aware has been designed. The MSA_BG algorithm is iterative and is based on the concept of Artificial Bee Colony metaheuristics and the concept of algorithmic and architectural spaces correlation. The metaphor of the ABC metaheuristics has been constructed and the functionalities of the agents has been defined. The conceptual parallel computational model has been designed. The algorithmic framework of the designed parallel algorithm has been constructed. MSA_BG algorithm has a hierarchical structure, which enables observing the principle of locality (independent calculations), and very high scalability (hives and swarms), so it is expected high efficiency implementations for petaflops supercomputers. Parallel performance evaluation and profiling of multiple sequence alignment on the basis of MSA_BG algorithm heterogeneous distributed high-performance computation resources have been proposed in this paper. The case study is investigating influenza virus nucleotide sequences and finding out consensus motifs and variable domains in the different segments. Parallel performance parameters, such execution time and acceleration, have been estimated experimentally. The performance estimation analyses have shown that the parallel system is well balanced both in respect to the workload and machine size.

Г.8.12. Borovska P., **Gancheva V.**, Georgiev I. Optimization of Multiple Sequence Alignment Algorithm ClustalW Using OpenMP and Vector Processing, Proceedings of the 11th International Conference on Challenges in Higher Education and Research in the 21st Century, 2013, Sozopol, Bulgaria, ISBN 978-954-580-325-3, pp. 181-184, https://elfe.tu-sofia.bg/cher21/index.php?nact=5021

Biocomputing and molecular biology are areas, demanding knowledge and skills for acquisition, storing, management, analysis, interpretation and dissemination of biological information. This requires the utilization of high performance computers and innovative software tools for the management of the vast information, as well as the deployment of innovative algorithmic techniques for the analysis and interpretation of data in order to get to the insight of the design and validation of life-science experiments. In this paper, we have performed investigation of comparative analyses of influenza virus nucleotide sequences on the basis of parallel computer simulation. For this purpose, a parallel multithreaded

computational model based on ClustalW algorithm for multiple sequence alignment is suggested and verified on the basis of parallel program implementation on a heterogeneous compact cluster.

Г.8.13. Borovska P., **Gancheva V.**, Asenov E., Georgiev I. Computational Aspects of In-silico Experiments for Investigating the Impact of the Host Genome on the Influenza Virus A Variability, *Journal Information Technologies and Control*, Vol.10, No 2/2012, pp. 8-14, ISSN 1312-2622, http://www.acad.bg/rismim/itc/sub/archiv/no2_2012.htm

Nowadays the study of the variability of influenza virus is a problem of very great importance. Influenza type A viruses cause epidemics and pandemics. The problem of restricting the spreading of pandemics and the treatment of the people infected by the influenza virus is widely based on the latest achievements of molecular biology, bioinformatics and biocomputing, as well as many other advanced areas of science. In silico biological sequence processing is a key for molecular biology. This scientific area requires powerful computing resources for exploring large sets of biological data. The paper presents parallel computational simulations for the case study of investigating the role of the host genome in the evolution and fast changeability of the influenza virus A on supercomputer BlueGene/P. The experimental framework is based on all available existing influenza virus A nucleotide sequences, the clustalw algorithm for multiple sequence alignment, the blast algorithm for sequence searching, the Philip software for philogenetic tree reconstruction and the recombination analysis tool for finding hot-spots of mutation/recombination in influenza A virus genomes.

Г.8.14. Borovska P., **Gancheva V.**, Tsvetanov S. Optimization and Scaling of Multiple Sequence Alignment Software ClustalW on Intel Xeon Phi, PRACE White Paper, 2014, Available online at https://prace-ri.eu/wp-content/uploads/wp138.pdf

This work is aimed to investigate and to improve the performance of multiple sequence alignment software ClustalW on the test platform EURORA at CINECA, for the case study of the influenza virus sequences. The objective is code optimization, porting, scaling and performance evaluation of parallel multiple sequence alignment software ClustalW for Intel Xeon Phi (the MIC architecture). For this purpose a parallel multithreaded optimization including OpenMP has been implemented and verified. The experimental results show that the hybrid parallel implementation utilizing MPI and OpenMP provides considerably better performance than the original code.

Г.8.15. Borovska P., **Gancheva V.**, Landzhev N. Code Optimization and Scalability Testing of an Artificial Bee Colony Based Software for Massively Parallel Multiple Sequence Alignment on the Intel MIC Architecture, PRACE White Paper, 2014, Available online at https://prace-ri.eu/wp-content/uploads/wp137.pdf

This activity with the project is aimed to investigate and to improve the performance of the multiple sequence alignment software MSA_BG on the computer system EURORA at CINECA, for the case study of the influenza virus sequences. The objective is code optimization, porting, scaling and performance evaluation of the parallel multiple sequence alignment software MSA_BG for Intel Xeon Phi (the MIC architecture). For this purpose a parallel multithreaded optimization including OpenMP has been implemented and verified. The experimental results show that the hybrid parallel implementation utilizing MPI and OpenMP provides considerably better performance than the original code.

Г.8.16. Charalampidou A., Daoglou P., Folias D., Borovska P., **Gancheva V.** A Hybrid Implementation of Massively Parallel Multiple Sequence Alignment Method Based on Artificial Bee Colony Algorithm, PRACE White Paper, 2014, Available online at https://prace-ri.eu/wp-content/uploads/wp127.pdf

The project focuses on performance investigation and improvement of multiple biological sequence alignment software MSA_BG on the BlueGene/Q supercomputer JUQUEEN. For this purpose, scientific experiments in the area of bioinformatics have been carried out, using as case study influenza virus sequences. The objectives of the project are code optimization, porting, scaling, profiling and performance evaluation of MSA_BG software. To this end we have developed hybrid MPI/OpenMP parallelization on the top of the MPI only code and we showcase the advantages of this approach through the results of benchmark tests that were performed on JUQUEEN. The experimental results show that the hybrid parallel implementation provides considerably better performance than the original code.

Г.8.17. Borovska P., **Gancheva V.** Massively Parallel Algorithm for Multiple Sequence Alignment Based on Artificial Bee Colony, PRACE White Paper, 2013, Available online at https://prace-ri.eu/wp-content/uploads/wp114.pdf

This activity with the project PRACE-2IP is aimed to investigate and improve the performance of multiple sequence alignment software ClustalW on the supercomputer BlueGene/Q, so-called JUQUEEN, for the case study of the influenza virus sequences. Porting, tuning, profiling, and scaling of this code has been accomplished in this aspect. A parallel I/O interface has been designed for effcient sequence dataset input, in which sub-groups' local masters take care of read operation and broadcast the dataset to their slaves. The optimal group size has been investigated and the effects of read buffer size on read performance has been experimented. The application to ClustalW software shows that the current implementation with parallel I/O provides considerably better performance than the original code in view of I/O segment, leading up to 6.8 times speed-up for inputting dataset in case of using 8192 JUQUEEN cores.

Г.8.18. Ko S.-H., Borovska P., **Gancheva V.** Optimization of Multiple Sequence Alignment Software ClustalW, PRACE White Paper, 2013, Available online at https://prace-ri.eu/wp-content/uploads/wp71.pdf

In silico biological sequence processing is a key task in molecular biology. This scientific area requires powerful computing resources for exploring large sets of biological data. Parallel *in silico* simulations based on methods and algorithms for analysis of biological data using high-performance distributed computing is essential for accelerating the research and reducing the investment. Multiple sequence alignment is a widely used method for biological sequence processing. The goal of this method is DNA and protein sequences alignment. This paper presents an innovative parallel algorithm MSA_BG for multiple alignment of biological sequences that is highly scalable and locality aware. The MSA_BG algorithm we describe is iterative and is based on the concept of Artificial Bee Colony metaheuristics and the concept of algorithmic and architectural spaces correlation. The metaphor of the ABC metaheuristics has been constructed and the functionalities of the agents have been defined. The conceptual parallel model of computation has been designed and the algorithmic framework of the designed parallel algorithm constructed. Experimental simulations on the basis of parallel implementation of MSA_BG algorithm for multiple sequences alignment on heterogeneouc compact computer cluster and supercomputer BlueGene/P have been carried out for the case study of the influenza virus variability investigation. The performance estimation and profiling analyses have shown that the parallel system is well balanced both in respect to the workload and machine size.

Scientific publications with IF/SJR

3.31.1.Sharabov M., Tsochev G., **Gancheva V.**, Tasheva A. Filtering and Detection of Real-Time Spam Mail Based on a Bayesian Approach in University Networks. *Electronics*. 2024; 13(2):374. . **IF=2.9 (2022) / SJR=0.644 (2023) / Q2, Scopus / WoS**

With the advent of digital technologies as an integral part of today's everyday life, the risk of information security breaches is increasing. Email spam, commonly known as junk email, continues to pose a significant challenge in the digital realm, inundating inboxes with unsolicited and often irrelevant messages. This relentless influx of spam not only disrupts user productivity but also raises security concerns, as it frequently serves as a vehicle for phishing attempts, malware distribution, and other cyber threats. The prevalence of spam is fueled by its low-cost dissemination and its ability to reach a wide audience, exploiting vulnerabilities in email systems. This paper marks the inception of an in-depth investigation into the viability and potential implementation of a robust spam filtering and prevention system tailored explicitly to university networks. With the escalating threat of email-based hacking attacks and the incessant deluge of spam, the need for a comprehensive and effective defense mechanism within academic institutions becomes increasingly imperative. In exploring potential solutions, this study delves into the applicability and efficacy of Bayesian filters, a class of probabilistic classifiers renowned for their aptitude in distinguishing between legitimate emails and spam messages. Bayesian filters utilize statistical algorithms to analyze email content, learning patterns and features to accurately categorize incoming emails. The results obtained from the Bayes approach exhibit positive outcomes, albeit falling short of reaching satisfactory levels. It is evident that substantial enhancements can significantly elevate the efficacy of our spam-filtering module through the implementation of various strategies. In summary, the experimental findings affirm the relevance of Bayes graphs in the realm of spam filtering. However, it is evident that substantial enhancements are imperative. As such, our future endeavors will concentrate on refining the milter itself, incorporating new data insights, and rigorously assessing the performance of alternative classifiers. These efforts are poised to furnish a definitive solution that transcends current limitations, establishing a robust and adaptive spam-filtering mechanism for optimal email security.

3.31.2. **Gancheva V.**, Galabova L. Platform for Learning and Virtual Reality in Animal Husbandry, *WSEAS Transactions on Information Science and Applications*, pp. 163-169, 2023, DOI 10.37394/23209.2023.20.19 **Scopus, SJR = 0.126 (2023), Q4**

The paper presents an integrated platform for open science and educational resource sharing, as well as an environment for distance learning and data analysis in animal husbandry derived from the learning resources of a given course in the system. The research is aimed at developing a computer-aided framework in the field of digitized education and creating new educational resources for distance learning in animal husbandry. The proposed platform provides a variety of ways to access and share educational resources through digital technologies and a repository to provide access to free online courses and learning content.

3.31.3. **Gancheva V.**, Georgiev I., Todorova V. X-Ray Images Analytics Algorithm based on Machine Learning, *WSEAS Transactions on Information Science and Applications*, 2023, pp. 136-145, DOI: 10.37394/23209.2023.20.16, **Scopus, SJR = 0.126 (2023), Q4**

The rapid development of information technology has led to a huge amount of data generated by large or complex systems and devices. Applications in information technology, medicine, and many other fields

generate large volumes of data that challenge analysts. Data mining analysis finds application in areas where statistical and analytical methods and the models built through them are not sufficient. The paper discusses sources of medical data, use cases, and data analysis in medicine, as well as methods and algorithms for data analysis. The purpose and objectives of the study, presented in the paper are to propose an algorithm for processing X-Ray images based on tools and techniques from the field of machine learning. The preprocessing phase is concerned with image transformation, feature extraction, and the selection of training and testing datasets. Preprocessing data enables the processing of data that would not otherwise be appropriate by adjusting the data to the specifications established by each data retrieval procedure. Each feature is examined in the second stage to identify and classify any potential patterns. In the final stage, the most effective model to capture the pattern or behaviour of the data is chosen using a machine learning algorithm. The proposed algorithm is verified using publicly available X-Ray image datasets consisting of four classes: Normal, Lung Opacity, Pneumonia, and COVID-19. A medical image classification workflow was designed for verification. In the experimental workflow, five algorithms in the field of machine learning are determined and implemented: Logistic Regression, Naive Bayes, Random Forest, SVM, and Neural Network. In comparison to the outcomes of Random Forest, Logistic Regression, Naive Bayes, and SVM, the findings of the experimental analysis and results demonstrate that Neural Networks produce the greatest results, and these results can be taken to be the most dependable.

3.31.4. Aleksieva-Petrova A., Gancheva V., Petrov M. APTITUDE Framework for Learning Data Classification Based on Machine Learning, *International Journal of Circuits, Systems and Signal Processing*, Volume 14, 2020, https://doi.org/10.46300/9106.2020.14.51, SJR=0.156 (2020) Q4, Scopus

Learning analytics refers to the machine learning to provide predictions of learner success and prescriptions to learners and teachers. The main goal of paper is to proposed APTITUDE framework for learning data classification in order to achieve an adaptation and recommendations a course content or flow of course activities. This framework has applied model for student learning prediction based on machine learning. The five machine learning algorithms are used to provide learning data classification: random forest, Naïve Bayes, k-nearest neighbors, logistic regression and support vector machines. The real time learning and gaming analytics of big data produced by modern e-learning platforms and educational games, for a learner-centric adaptation of technology enhanced learning is one of main challenge. The paper proposed software architecture for adaptation and recommendation of course content and activities based on learning analytics. It helps to structure and storage of big data from heterogeneous sources as both LMS and educational game; identify patterns by analyzing learners' behavior and allowing data analyses with descriptive, predictive, and prescriptive results. The student learning prediction algorithm based on machine learning for processing and analysis of data and knowledge discovery with respect to main learner and teacher activities is designed. Experimental results are presented and discussed. The experimental data set is obtained from learning management system and contains of 63774 instances characterized by 7 attributes. Log files in Moodle system are used for analyses. For analysis are used clustering algorithms provided by Weka: Expectation Maximization, Hierarchical Clustering, Simple K-Means, and X-Means to find correlation in graded activities. The paper is implemented other five ML algorithms in order to validate their applicability for student learning prediction. Based on different analytical models which created after execution of the feature extraction and data set reduction process, the prototype will be validate and verification the usability of the proposed architecture.