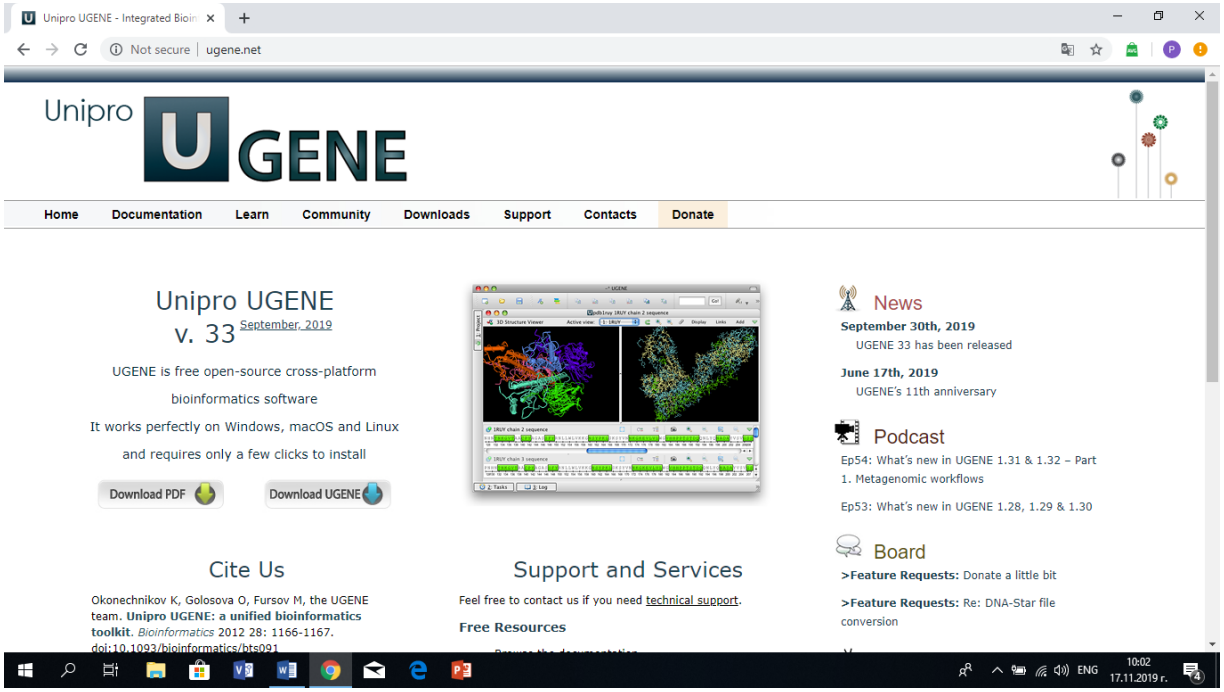
Биоинформатични платформи

Биоинформатичните платформи са специализирани за научни работни потоци за целите на биоинформатиката. Понастоящем, съществува широк спектър от биоинформатични платформи, като основавайки се на положителния опит от предходни изследвания, фокусът е върху 3 платформи – Unipro UGENE, Taverna и Galaxy.

Unipro UGENE представлява мултиплатформен софтуер с отворен код, чието основно предназначение е да подпомага молекулярните биолози в управлението на техните in silico експерименти, както и при визуализацията и анализа на експерименталните резултати. UGENE интегрира широк спектър биоинформатични софтуерни инструменти в рамките на общ потребителски интерфейс. Поддържат се множество биологични формати на данни и се осигурява извличането на данни от отдалечени източници. Съдържа модули за визуализация на биологични обекти като аннотирани секвенции на геноми, асемблирани NGS данни, множествено подравняване на биологични секвенции, конструиране на филогенетични дървета и тримерни структури на протеини.



Фиг. 1 Биоинформатична платформа Unipro UGENE

UGENE предлага среда за визуализация при създаването на многократно изпълними работни потоци, които могат да бъдат изпълнявани както на локални, така и

на отдалечени високо производителни изчислителни инфраструктури. UGENE е написан на езика за програмиране C++ в работната рамка Qt. Вградената plugin системи и структурираният приложен програмен интерфейс UGENE API осигурява възможности за разширяването на системата от софтуерни инструменти с нови функционалности.

Workflow Designer на UGENE е софтуерен инструмент за визуализация при изграждането на сложни аналитични конвейери. UGENE е самостоятелно приложение и не изисква инсталирането на допълнителни приложения.

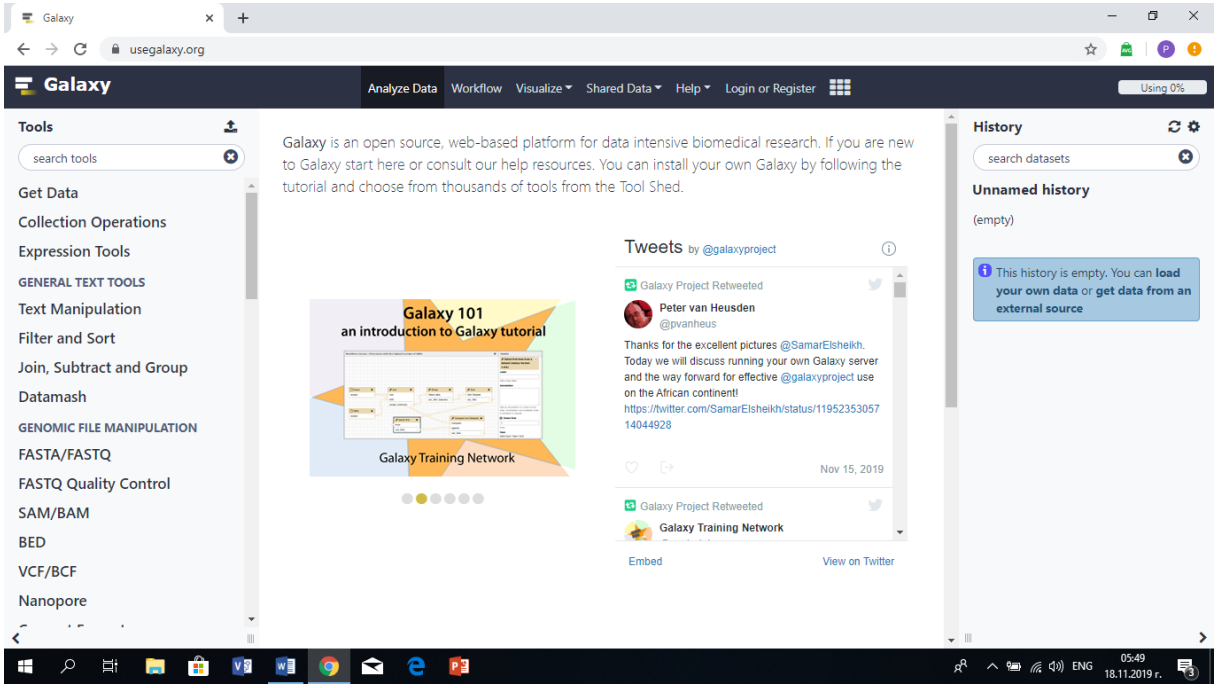
Системата за биоинформатични работни потоци ***Apache Taverna*** е система за управление на работни потоци с отворен код и независима от домейните. Представлява набор от софтуерни инструменти, използвани за проектиране и изпълнение на научни работни потоци за *in silico* експериментиране. *Taverna* е проект с отворен код от 2003 г. с участието на различни академични и индустриални партньори. През преиода 2014г. *Таверна* става проект в инкубатора на *Apache Software Foundation* и променя името си на *Apache Taverna* (инкубационен - *incubating*). Проектът разработва *Apache Taverna 3.x* като лицензът е променен от *LGPL 2.1* на *Apache License 2.0*. *Apache Incubator* е порталът за проекти с отворен код, предназначени да станат пълноценни проекти на *Apache Software Foundation*.

Проектът за инкубатора е създаден през октомври 2002 г., за да осигури подготовката на външни проекти за присъединяването им към проектите и базите данни на софтуерната фондация *Apache*. Всички дарения на код от външни организации и съществуващи външни проекти, които желаят да се преместят в Apache, трябва да влязат през Инкубатора.

Проектът за инкубатора на *Apache* от една страна служи като временен проект за контейнери, докато инкубационният проект бъде приет и се превърне в проект на най-високо ниво на софтуерната фондация Apache или стане подпроект на подходящ проект, като например Jakarta Project или Apache XML. От друга страна, проектът за инкубатора документира как работи Фондацията и как да се направят проектите в нейните рамки, което обхваща документирането, ролите и политиките в рамките на софтуерната фондация Apache и нейните членове.

Galaxy е уеб базирана платформа с отворен код за интензивни биомедицински научни изследвания. Представлява платформа за управление на научни работни потоци, а също така осигурява средства за интегриране на биологични данни. Създадена е от Galaxy Community (първа версия през 2005г.) и обхваща следните 3 проекта: Federated Galaxy Roadmap, Gen3 Galaxy integration, Remote Data Galaxy. Платформата е Linux

базирана и е написана на езиците Python и JavaScript. Galaxy първоначално е предназначена за анализ на биологични данни, по-специално геномика. Наборът от налични инструменти се разширява значително с годините и Galaxy вече се използва и за експресия на гени, асемблиране на геноми, протеомика, епигеномика, транскриптомика и множество други дисциплини в науките за живота. Самата платформа всъщност е агностичен домейн и може да се приложи на теория към всяка научна област. Например, създадени са Galaxy сървъри за анализ на изображения, изчислителна химия и дизайн на лекарства, космология, климатично моделиране, социални науки, и лингвистика.



Фиг. 2 Биоинформатична платформа Galaxy

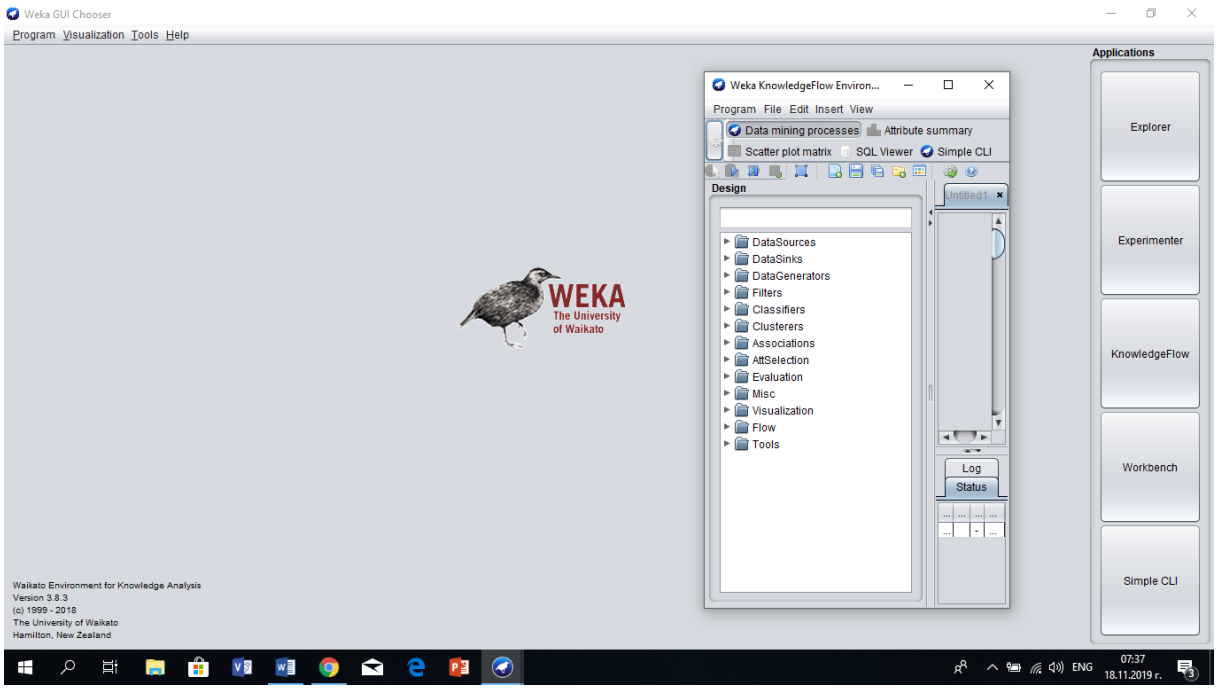
Galaxy поддържа прозрачността в научни изследвания, като дава възможност на изследователите да споделят някои от техните Galaxy обекти публично или с конкретни лица. Споделените елементи могат да бъдат разгледани подробно, да бъдат преизпълнени по желание и копирани и модифицирани, за да се тестват хипотези.

Обектите, които се поддържат от Galaxy са история, работни потоци, набори данни и страници. Историите, работните потоци и наборите данни могат да съдържат анотации на изследователя. Galaxy *Pages* осигуряват средства за създаването на виртуални научни доклади, описващи целия експеримент.

Работни рамки за машинно обучение и анализ на данни

С цел осигуряване на минимални финансови разходи за използване на интелигентното интегрирано дигитално решение за рака на гърдата фокусът е върху работни рамки за анализ на данни, които не са базирани на облачни услуги.

Weka (**W**aikato **E**nvironment for **K**nowledge **A**nalysis) е софтуер за машинно обучение, имплементиращ колекция от алгоритми за машинно обучение на Java на Университета Waikato в Нова Зеландия. Съдържа софтуерни инструменти за подготовка на данните, класификация, регресия, клъстериране, асоциативни правила и визуализация.

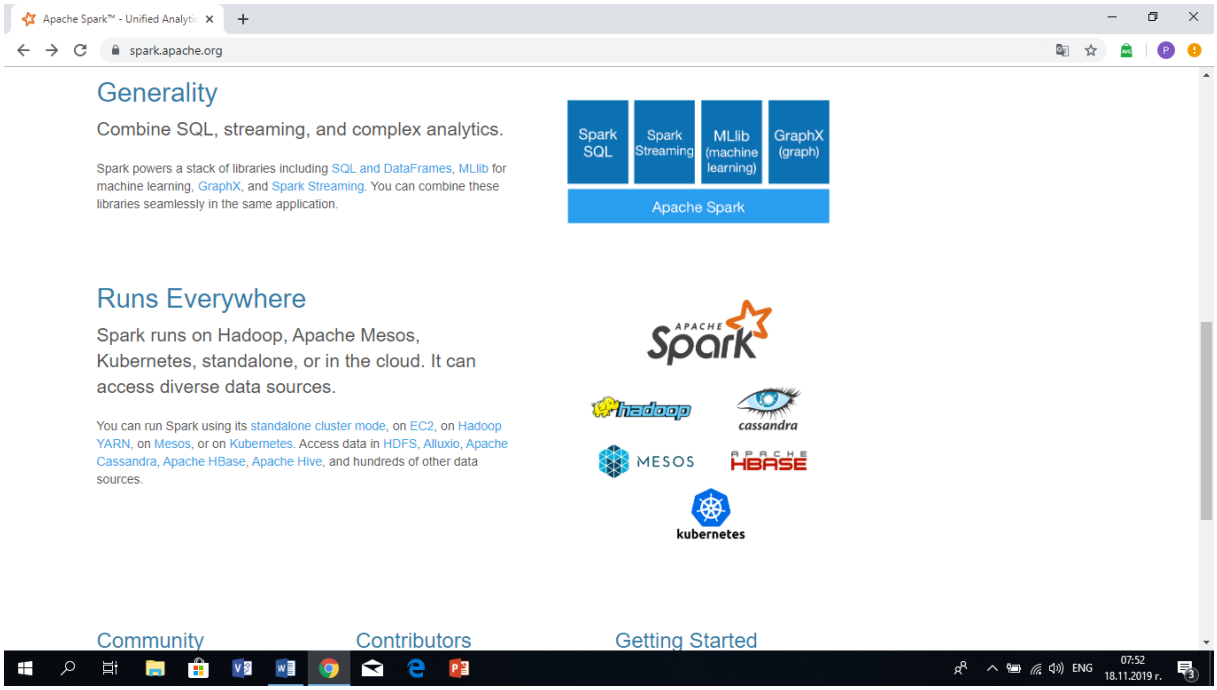


*Фиг. 3 **W**aikato **E**nvironment for **K**nowledge **A**nalysis (WEKA)*

Всички техники на Weka се основават на предположението, че данните са налични като един плосък файл или връзка, където всяка точка от данни се описва от фиксиран брой атрибути (обикновено числови или номинални атрибути, но някои други типове атрибути също се поддържат). Weka осигурява достъп до SQL бази данни с помощта на Java Database Connectivity и може да обработва резултата, върнат чрез заявка към база данни. Weka осигурява достъп до задълбочено обучение с Deeplearning4j. Не поддържа мултирелационно извличане на данни, но има отделен софтуер за преобразуване на

колекция от свързани таблици от бази данни в една таблица, която е подходяща за обработка с Weka. Друга важна област, която в момента не е обхванатата от алгоритмите, включени във Weka разпределението, е моделирането на последователности.

Apache Spark е работна рамка с отворен код за клъстерни изчисления, осигуряваща широк спектър от софтуерни инструменти за анализ на данни и откриване на знания. Apache Spark (unified analytics engine for large-scale data processing) е работна рамка с отворен код, поддържана от Apache Software Foundation. Осигурява приложни програмни интерфейси на високо ниво (APIs) за езиците за програмиране Scala, Java, Python, и R. Spark е платформа за разпределена обработка на големи масиви и потоци данни. Spark съдържа библиотеката MLlib, която е предназначена за анализ на големи масиви от данни с помощта на алгоритми за машинно обучение. Spark използа езиците за програмиране Scala, Python, Java, R, и SQL.



Фиг. 4 Apache Spark