ПРОЕКТИРАНЕ НА СИСТЕМА ЗА БИОИНФОРМАТИЧНА ОБРАБОТКА НА МЕТАГЕНОМНИ ДАННИ

Александър Кирилов

ИИКТ-БАН

Високопроизводителни пресмятания в полза на изследователите и обществото

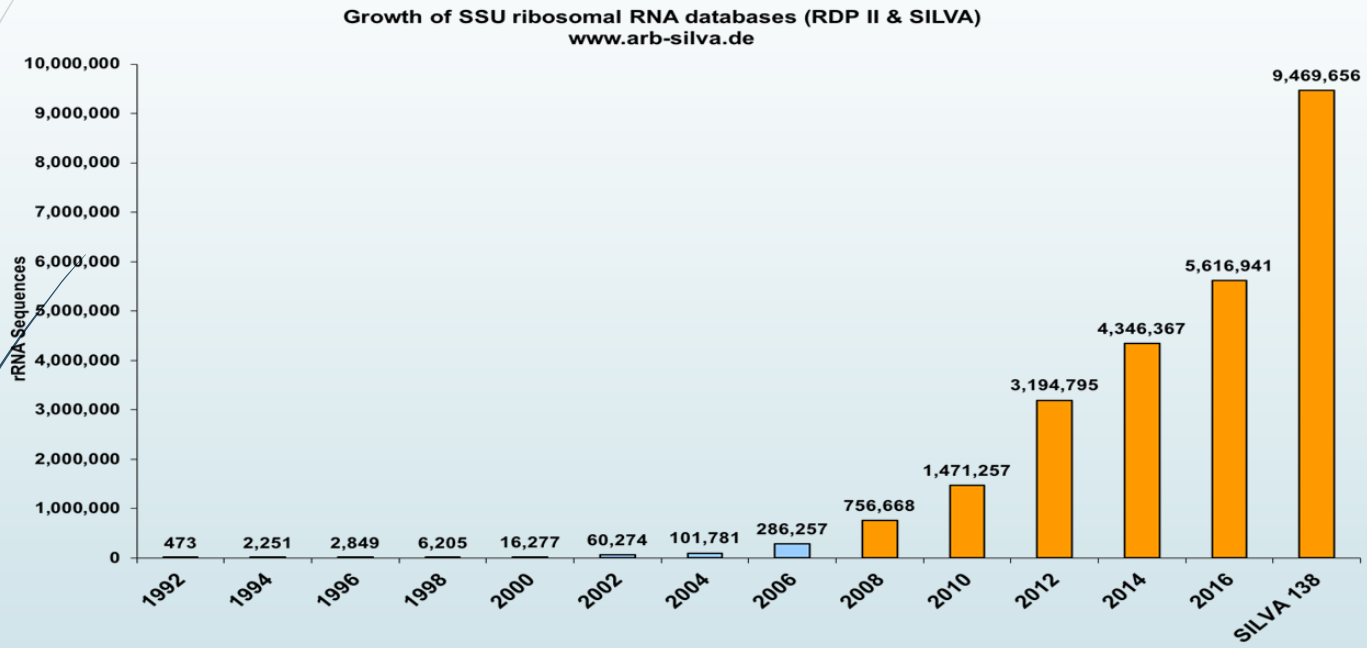
Съдържание

- Цели и задачи
- Биоинформатична обработка на данни
- Метагеномни изследвания
- Galaxy Software
- Реализация на сървърна система
- Модел за обработка на метагеномни данни
- Резултати
- Заключение

Основна цел

- Конфигуриране и изграждане на комплексна работеща сървърна система
- Инсталация на Galaxy software за изследвания в областта на биоинформатиката
- Тестове върху Amplicon метод с16S rRNA данни
- Метагеномни изследвания.

Биоинформатика данни



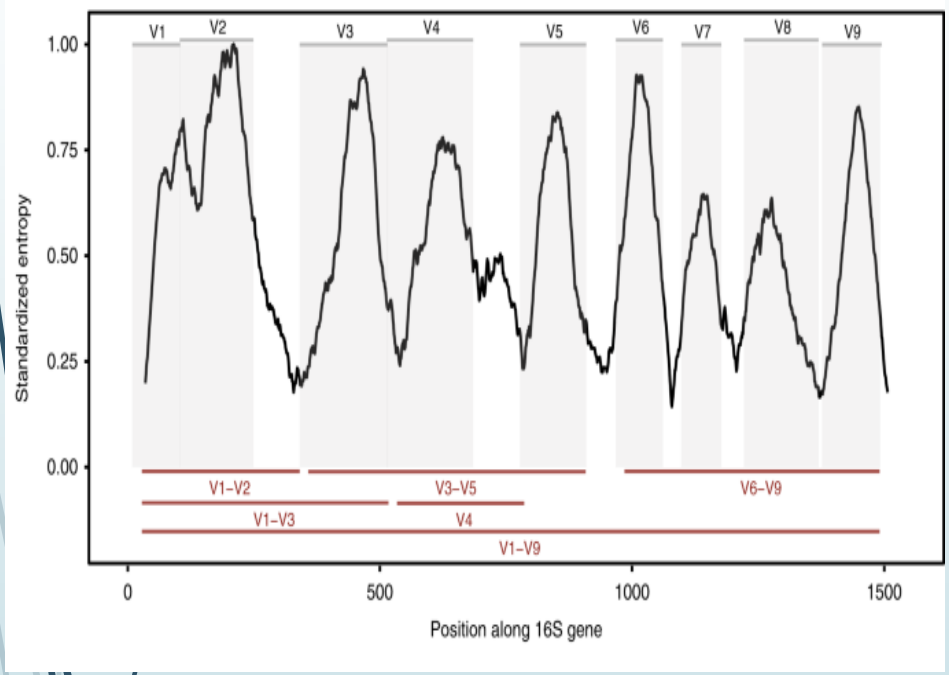
Фигура 1. Растеж на базите данни за рибозомна РНК в базата данни Silva

Приложение на метагеномиката

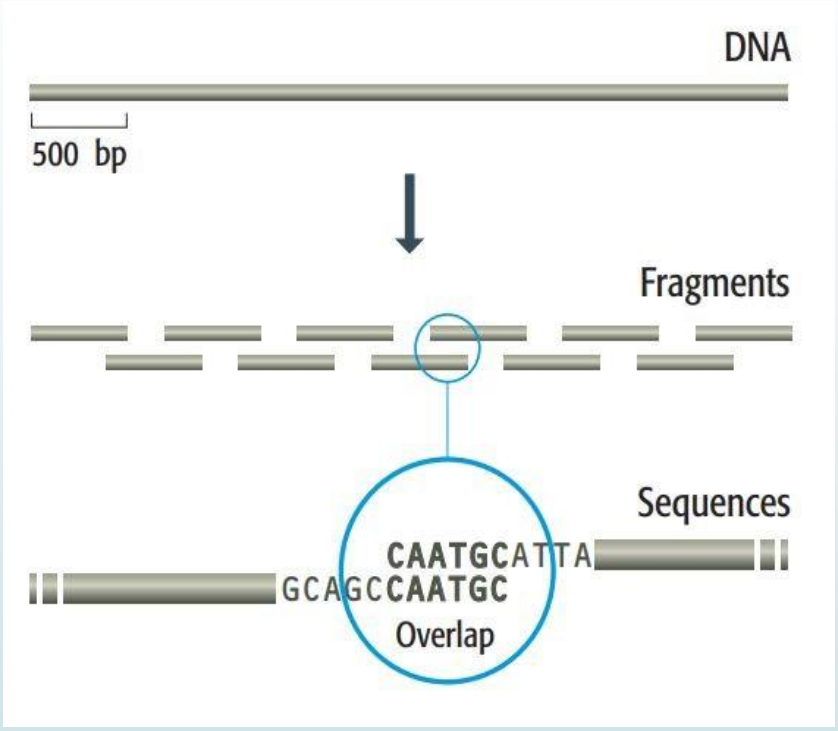
Отрасли на промишлеността	Приложение
Селско стопанство	Растеж на растения. Откриване на болести в културите и добитъка.
Биогориво	Индустриални приложения в производството на биогорива.
Биотехнология	Производства на лекарства. Фармацевтични продукти. Малацидиновите антибиотици.
Екология	Представа за екологичните общности. Обхват на инвазивните и застрашени видове. Проследяване на сезонни популации.
Диагностика на инфекциозни заболявания	Чувствителен и бърз метод за диагностициране на инфекция чрез сравняване на генетичен материал.

Таблица 1. Области на приложение на метагеномиката

Метагеномика - методи



Фигура 2. Amplicon метод. Променливи региони (V1-V9) на 16S rRNA ген.



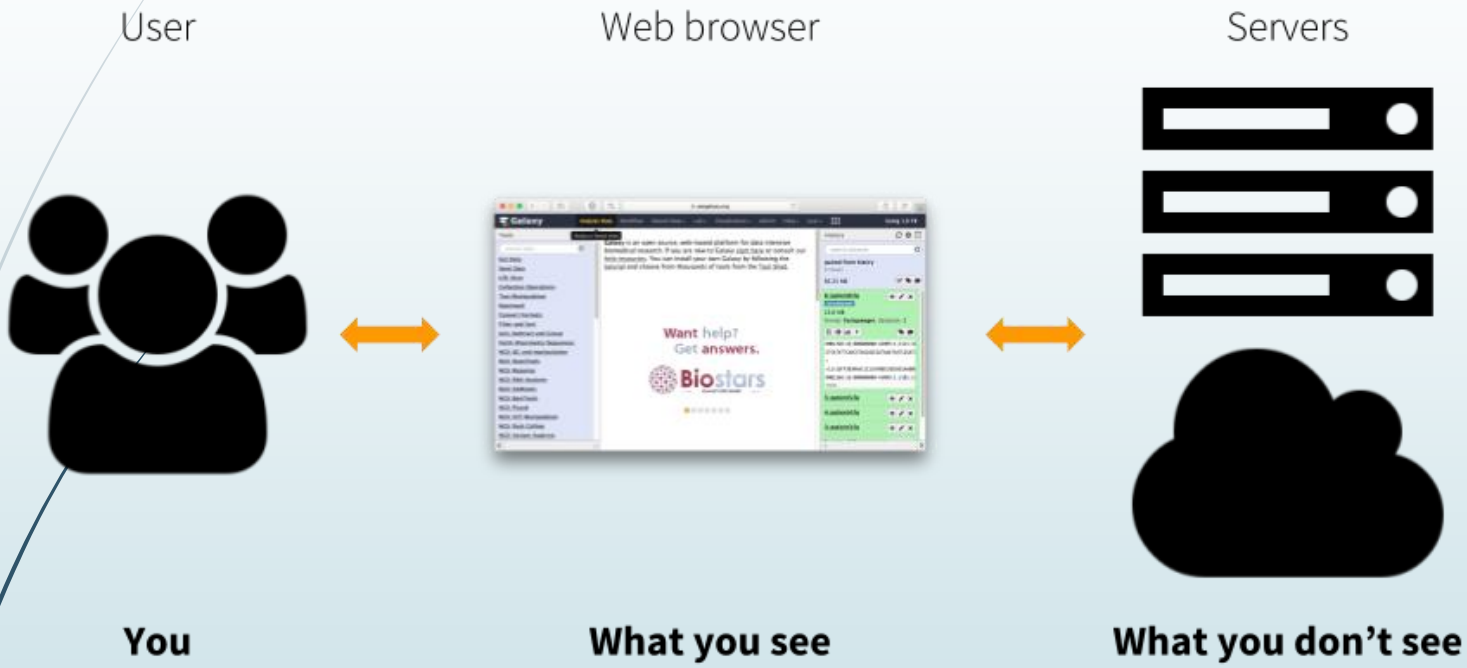
Фигура 3. Процес на изпълнение на секвенирането чрез "Shotgun" метод.

Galaxy Software

Предимства:

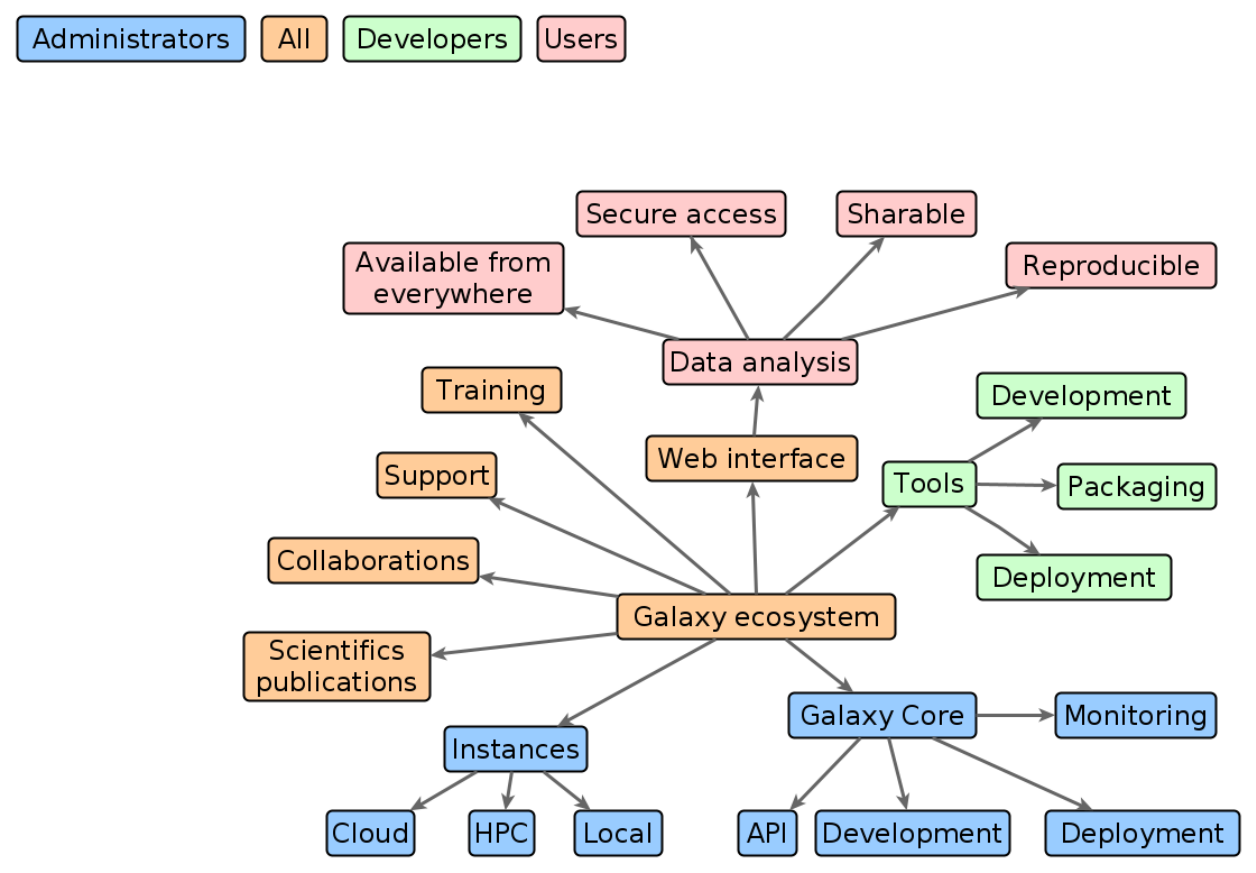
- Сайт за достъп - https://usegalaxy.eu/
- Отворения код и софтуерните пакети за анализ;
- Използване на голям набор от данни в биоинформатиката;
- 8500 готови инструменти, готови за извършване на различни видове манипулации;
- Голям набор от данни, готови обучения и разработени ръководства за анализиране;
- Безплатен достъп до ресурсите й;
- Създаването на работни потоци за анализ на данни;
- Създава общност от потребители, учени, изследователи, които обменят опит в областта.

Galaxy Software



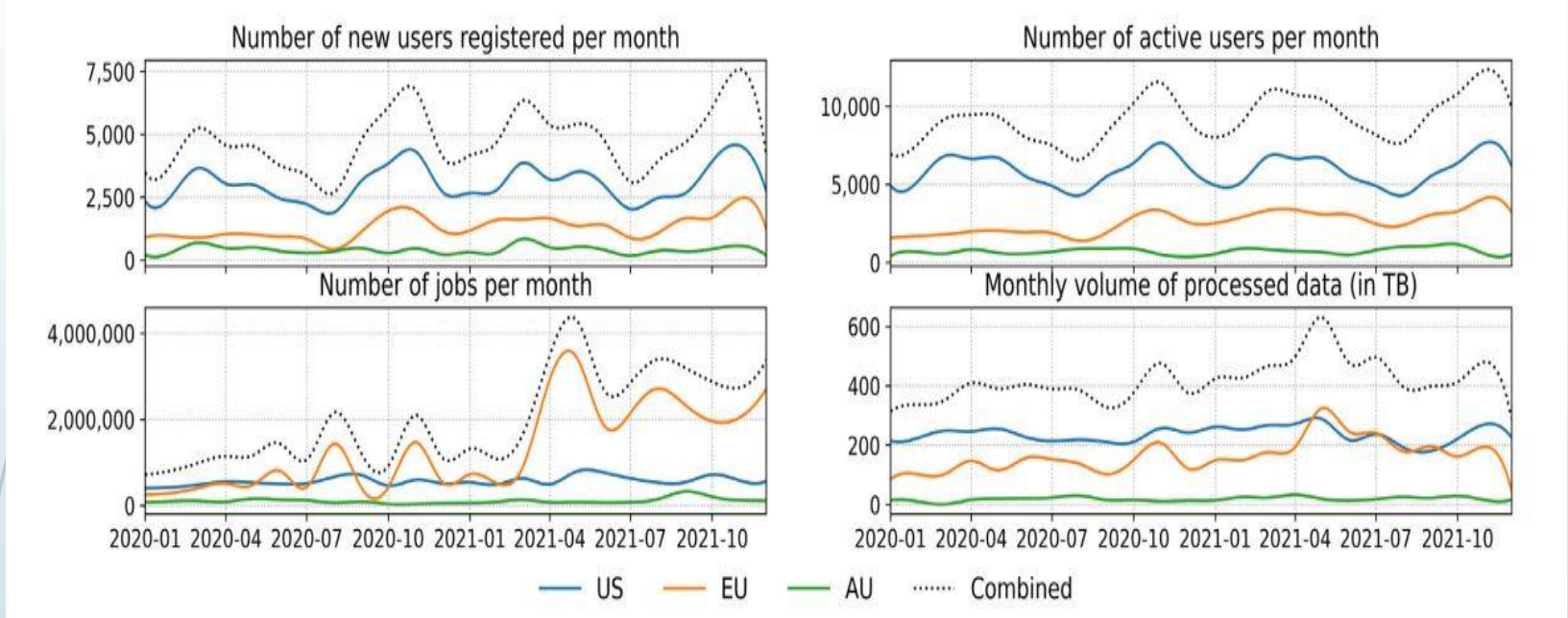
Фигура 4. Взаимодействието на потребителите с Galaxy Europe платформата

Galaxy Software



Фигура 5. Взаимодействието на потребителите с Galaxy Europe платформата

Galaxy Software



Фигура 6. Графики на използването на безплатни услуги от страна на изследователите, броят на регистрираните потребители и използването на ресурсите

Реализация на сървърна система

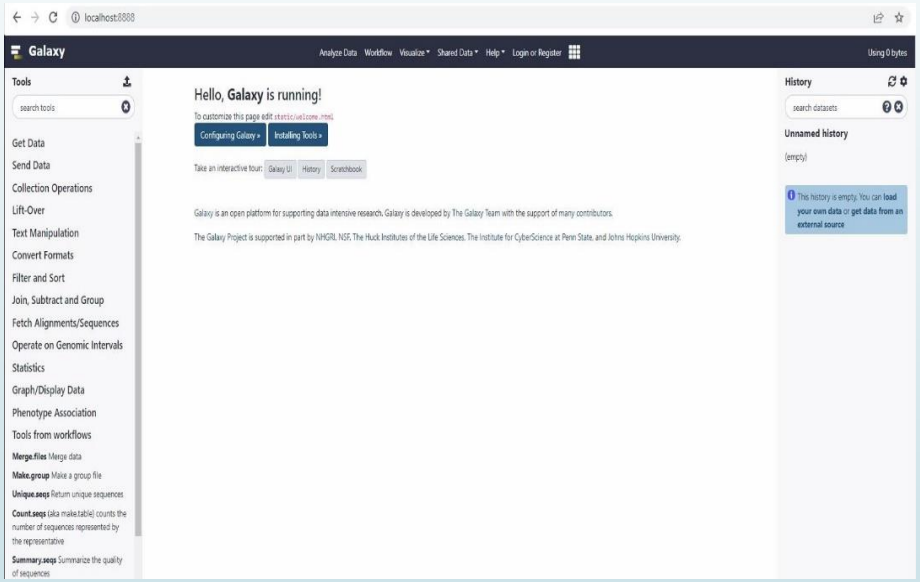
- 16 процесорни ядра (Cores)
- 4.8 TB дисково пространство за съхранение на данни, както и възможност за добавяне на още дискове с памет
- 32 GB операционна памет, със слотове за допълнителна памет
- Ubuntu OS

Инсталация на Galaxy Europe върху сървърна система

- Инсталация с Docker (https://www.docker.com/)

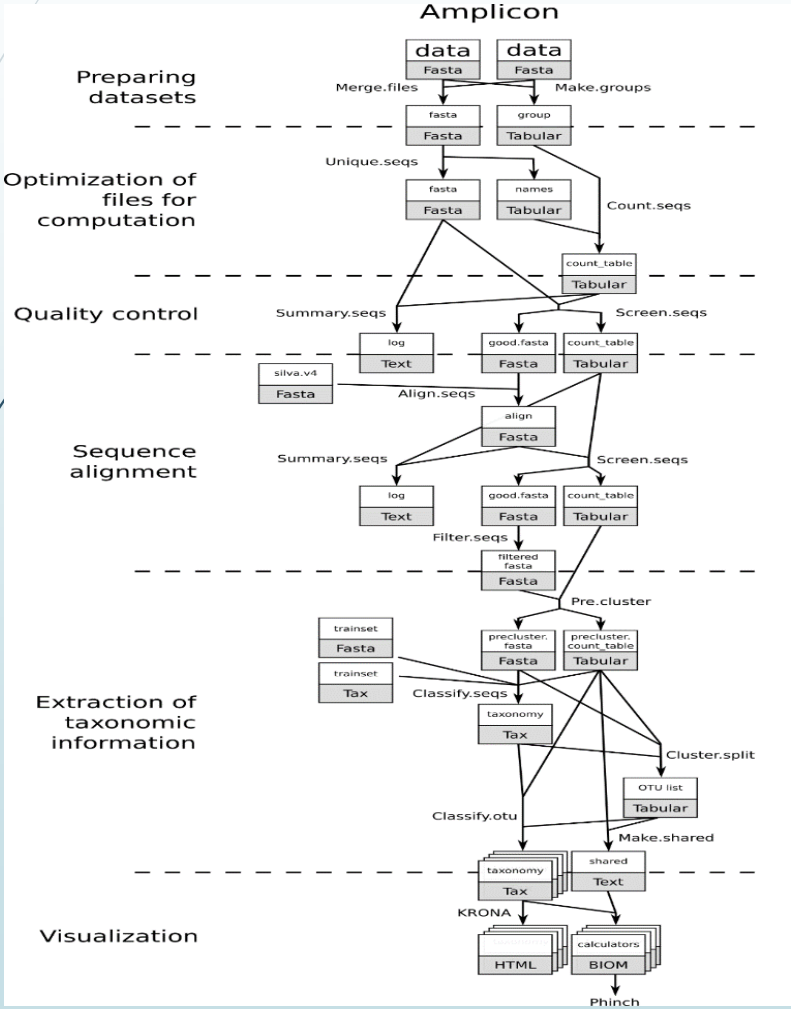
```
root@agalaxy01:~# ./install.sh
+ sh -c DEBIAN_FRONTEND=noninteractive apt-get install -y -qq docker-ce-rootless-
extras >/dev/null
+ sh -c docker version
Client: Docker Engine - Community
 Version:           20.10.17
 API version:       1.41
 Go version:        go1.17.11
 Git commit:        100c701
 Built:             Mon Jun  6 23:02:46 2022
 OS/Arch:           linux/amd64
 Context:           default
 Experimental:      true

Server: Docker Engine - Community
 Engine:
  Version:          20.10.17
  API version:      1.41 (minimum version 1.12)
  Go version:       go1.17.11
  Git commit:       a89b842
  Built:            Mon Jun  6 23:00:51 2022
  OS/Arch:          linux/amd64
  Experimental:     false
 containerd:
  Version:          1.6.7
  GitCommit:        0197261a30bf81f1ee8e6a4dd2dea0ef95d67ccb
 runc:
  Version:          1.1.3
  GitCommit:        v1.1.3-0-g6724737
 docker-init:
  Version:          0.19.0
  GitCommit:        de40ad0

================================================================================
```

Фигура 7. Инсталираният Docker клиенти и сървър



Фигура 8. Начална страница на Galaxy Europe софтуер на локалната сървърна платформа

Модел за обработка на метагеномни данни Amplicon



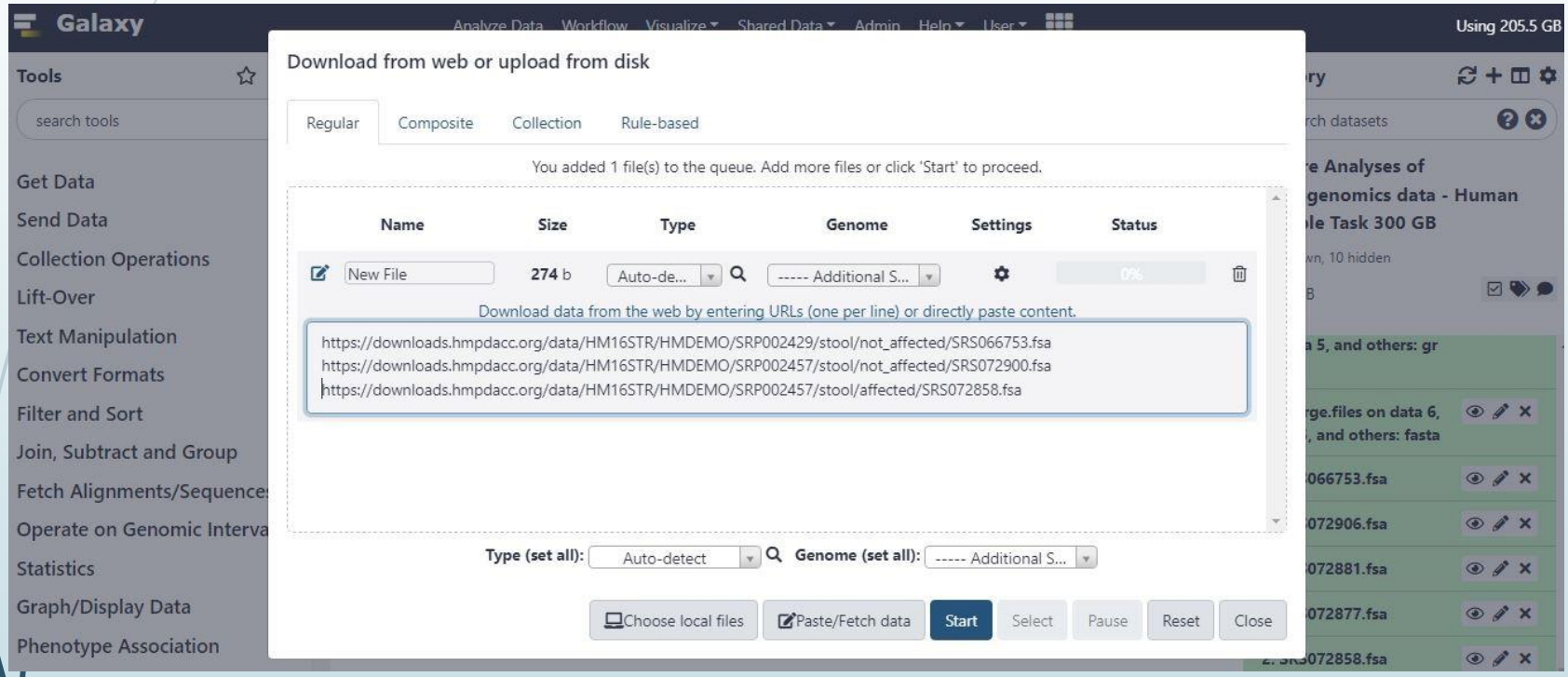
Фигура 9. Структурен модел на Amplicon анализ на метагеномни данни

Метагеномни данни, използвани при тестовете за производителност (1/3)

Source	Dataset	Seqs	Size(MB)
HMP Ileum	Test_300	527135	300
HMP Ileum	Test_700	1121140	700
HMP Ileum	Test_1500	2577144	1500

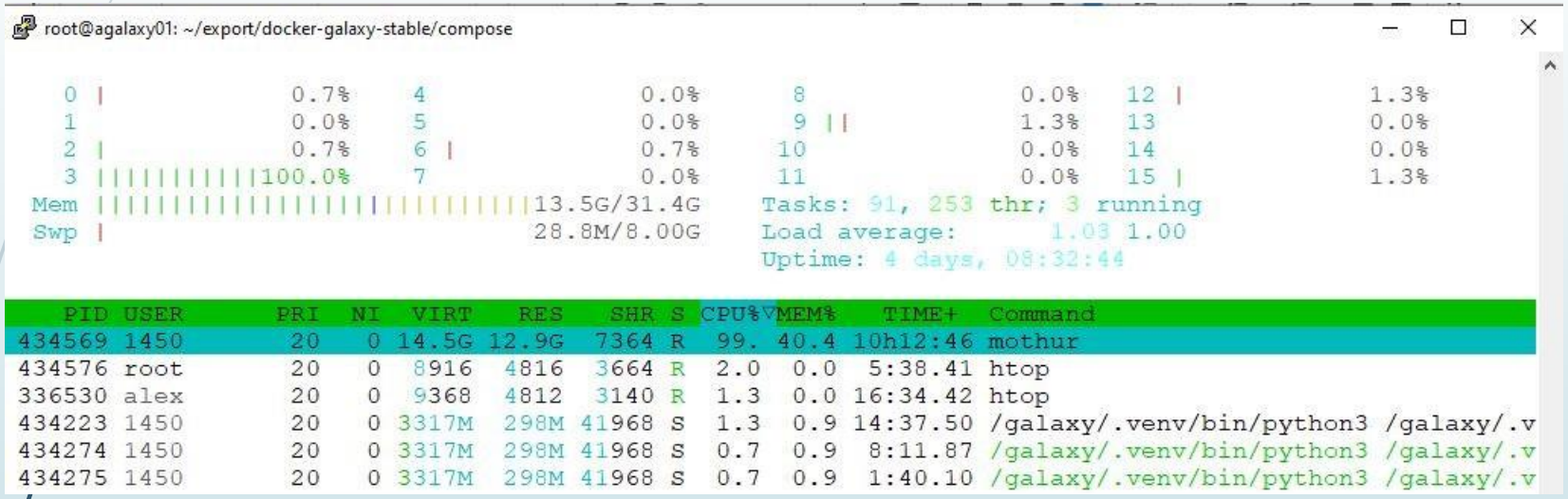
Таблица 2. Тестови данни: Брой на последователности (Seqs); Големина (Size)

Метагеномни данни, използвани при тестовете за производителност (2/3)



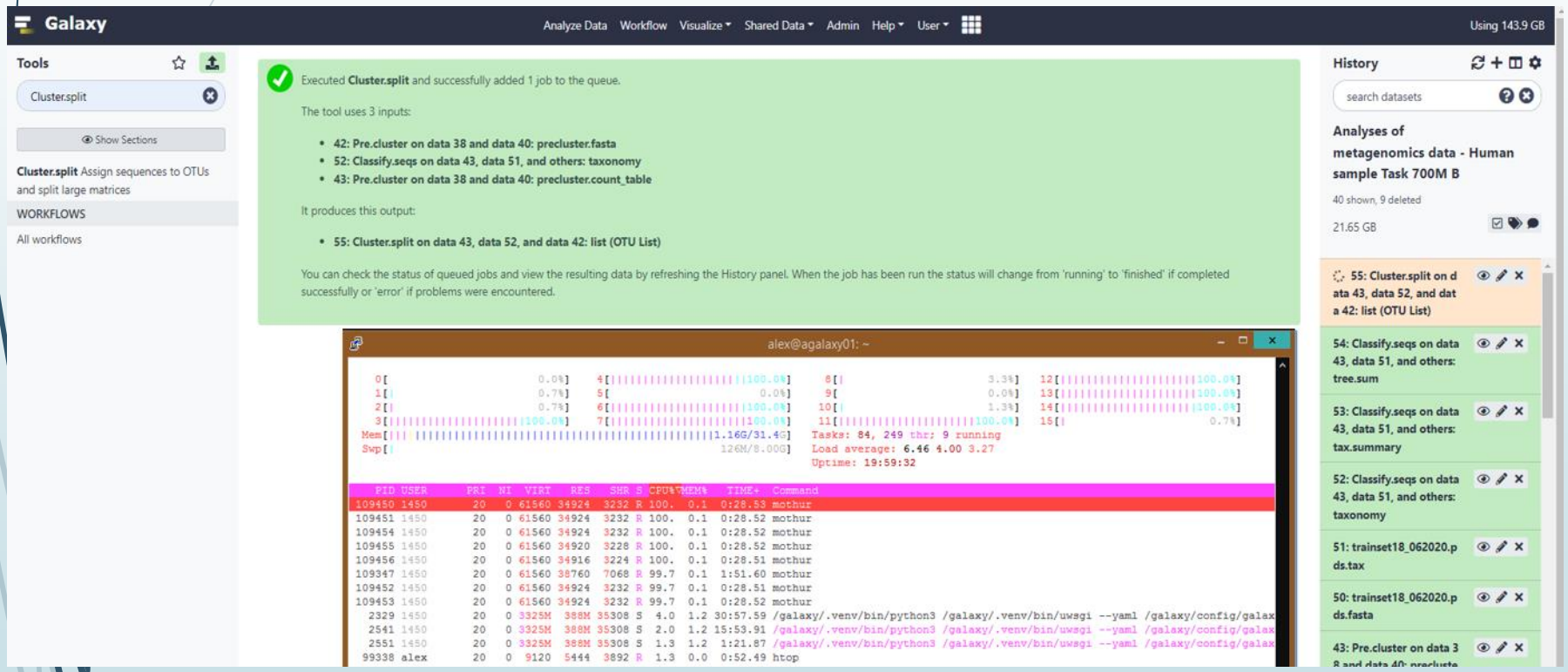
Фигура 10. Зареждане на данните и стартиране на процеса.

Метагеномни данни, използвани при тестовете за производителност (2/3)



Фигура 11. Натоварване на системата. Задачата се изчислява над 10 часа и използва 1 ядро и 13 GB RAM памет.

Метагеномни данни, използвани при тестовете за производителност (3/3)

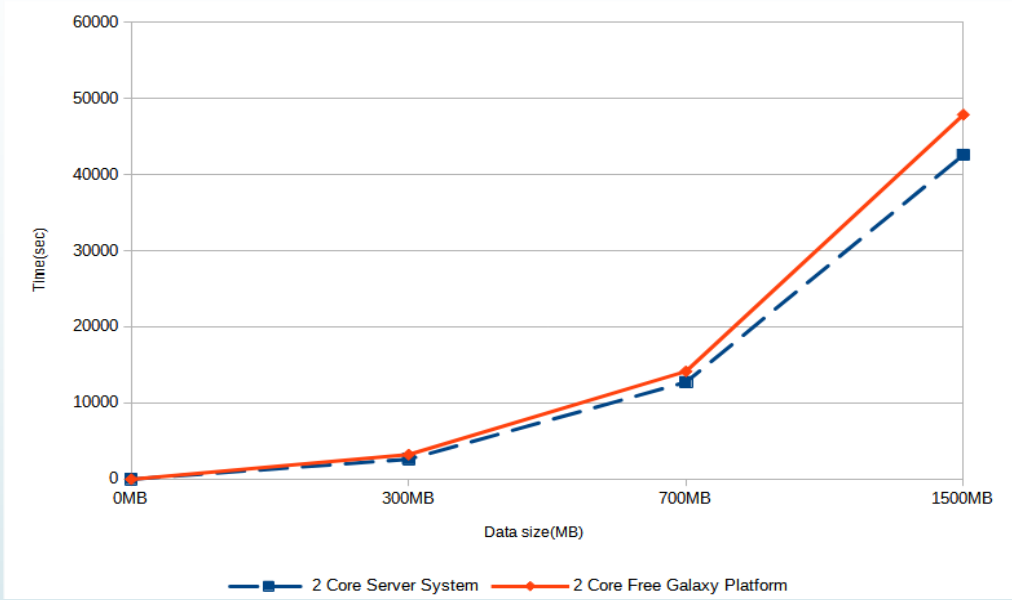


Фигура 12. Натоварване на локална сървърна системата с 8 процесорни ядра при изчисление.

Резултати (1/4)

Dataset Size (MB)	Time (sec)	
	2 Core server system	2 Core Free Galaxy Platform
300	2618	3223
700	12724	14145
1500	42550	47833

Таблица 3. Времена за изчисление на двете системи с 2 ядра

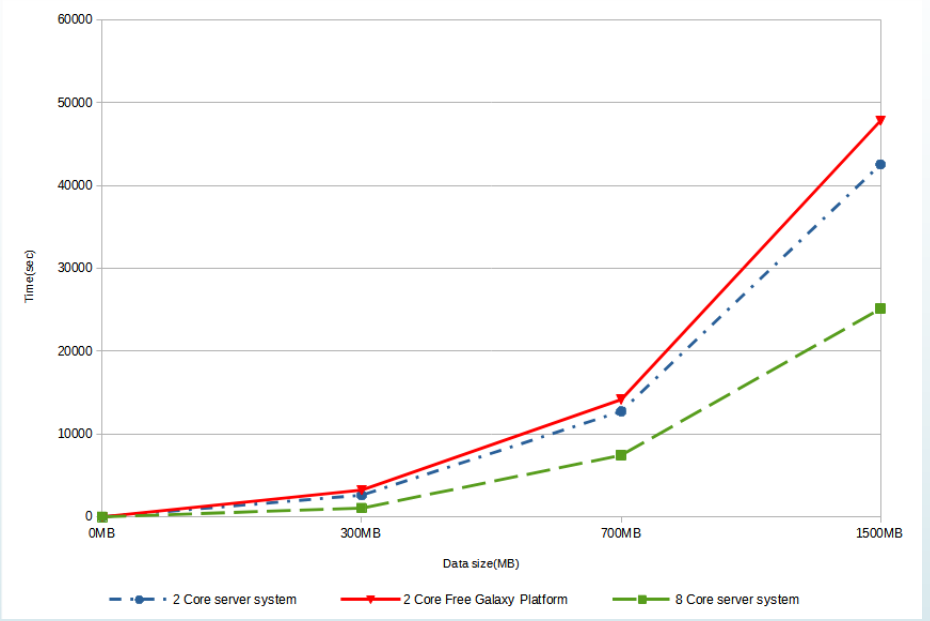


Фигура 13. Графично разпределение на времената за изчисление при различна големина на данните с еднаква конфигурация на системите

Резултати (2/4)

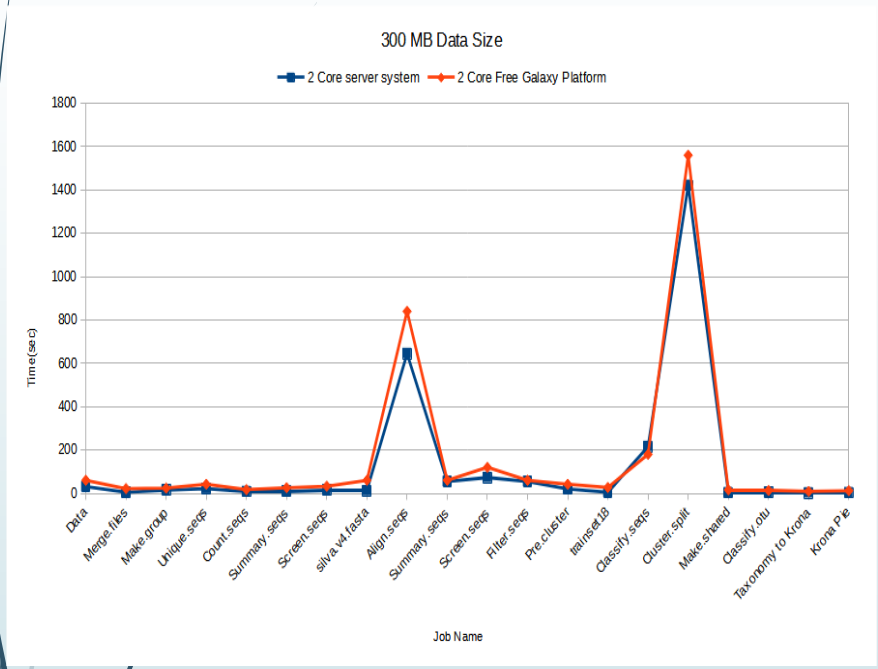
Dataset Size (MB)	Time (sec)		
	2 Core server system	2 Core Free Galaxy Platform	8 Core server system
300	2618	3223	1053
700	12724	14145	7433
1500	42550	47833	25112

Таблица 4. Времена за изчисление на три системи

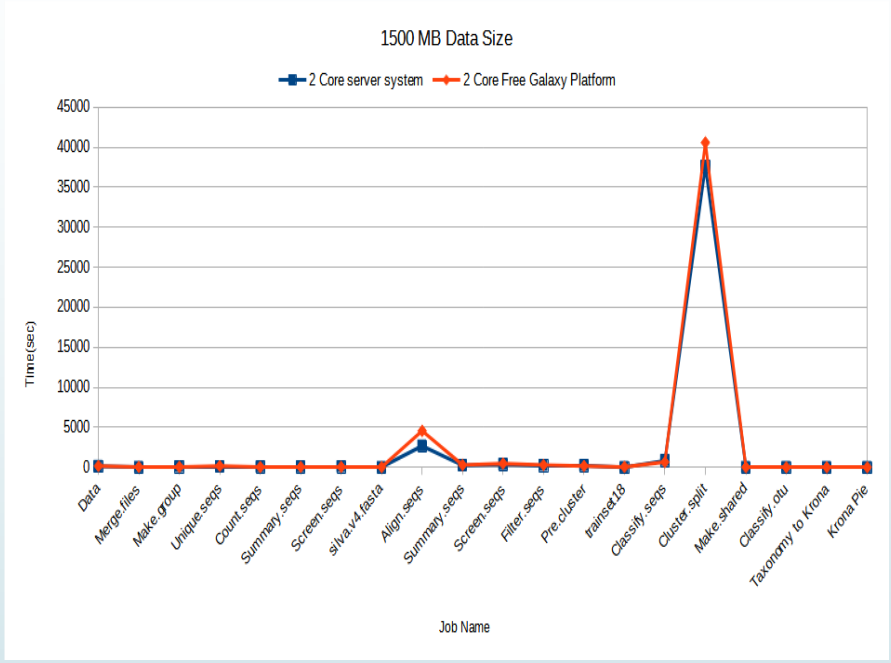


Фигура 14. Графично разпределение на времената за изчисление при различна големина на данните при различни конфигурации на системите

Резултати (3/4)

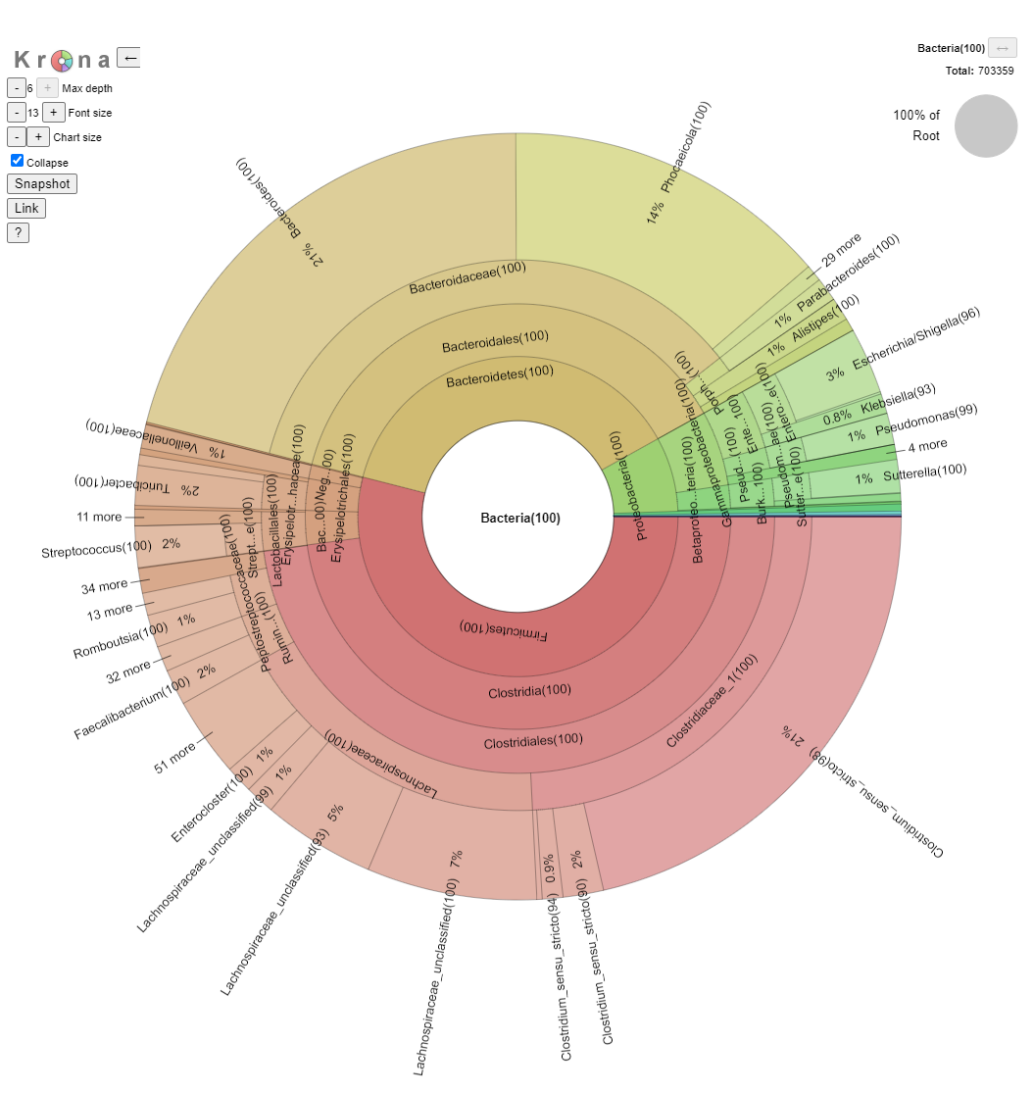


Фигура 15. Сравнителна графика за времето за изпълнение на първи експеримент (300 MB) върху двата вида системи



Фигура 16. Сравнителна графика за времето за изпълнение на първи експеримент (1500 MB) върху двата вида системи

Резултати (4/4)



Фигура 17. Krona Pie HTML интерактивно графично представяне на резултат от изследване върху данни с големина 1500 MB.

Заключение

- *Galaxy Software* е удобен. Различните му инструменти осигуряват бърз достъп до данните, осигуряващо систематизирани работни потоци за анализ и обработка.
- Разработих и конфигурирах сървърна система за нашите нужди като изследователи, за да намалим времето за обработка на метагеномни данни.
- Съобразена е структурата и големината на данните, които са обработени и анализирани.
- Извършените експерименти демонстрират, че съществува значителна връзка между характеристиките на данните за входната последователност и изчислителното време, необходимо за обработка на тези данни.
- При обработката на големи бази данни е необходимо използването на паралелни пресмятания върху компютърни клъстери или суперкомпютри. Това ще намали времето за обработка и изследвания значително.

Благодаря за вниманието!