Софийски университет „Св. Климент Охридски”
Факултет по Математика и Информатика
Катедра „Информационни технологии”
Специализация „Био- и медицинска информатика”

Дипломна работа
Тема:

Информационна система за анализ на генна експресия

Дипломант:

Златомир Златомиров Златев, ФN№ M22147

Ръководител:

проф. Иван Иванов,

Department of Physiology and Pharmacology,

Texas A&M University

Консултант:

доц. Д-р Антоний Попов,

катедра “Информационни технологии”,

СУ “Св. Климент Охридски”

Юли 2009 г
гр. София

Съдържание

Структура на дипломната работа 5
1. Въведение 6
1.1. Генна експресия 6
1.2. кДНК матрица (cDNA microarray) 6
1.3. Анализ на генна експресия 6
1.4. Мотивация 7
1.5. Цел и задачи на дипломната работа 7
2. Биологична основа. ДНК. Гени. Генна експресия. Ниво на генната експресия. кДНК матрици. Стандартизация. 9
3. Информационна система за анализ на генна експресия "zRMicroArray – Microarray Analysis tool" 14
3.1. Използвани програмни средства и дизайн на софтуера 15
3.2. Превод на различни езици 18
3.3. Информационен екран 18
3.4. Меню "Проект" 19
3.5. Данни 20
3.5.1. Импортиране 20
3.5.2. Основни статистически показатели 21
3.5.3. Разпределение на данните и филтриране на ниво файл 22
3.5.4. Насоки за развитие 23
3.6. Групиране на данните от експеримента 23
3.7. Предварителен анализ 25
3.7.1. Основни статистически показатели 25

3.7.2. Тест за нормално разпределение на гените в различните групи. Генериране на липсващи стойности. 26
3.7.3. Насоки за развитие 27
3.8. Филтри 27
3.8.1. Контроли 29
3.8.2. Флагове 30
3.8.3. Насоки за развитие 30
3.9. Трансформации 31
3.9.1. Реализирани трансформации 31
3.9.2. Насоки за развитие 32
3.10. Статистически анализ 32
3.10.1. Тест за нормално разпределение на гените в различните групи. Възможни комбинации от групи. 32
3.10.2. Дисперсионен анализ (ANOVA) 34
3.10.3. Погрешност при идентифициране (False Discovery rate) 35
3.10.4. Насоки за развитие 35
3.11. Сравняване на анализи от един и същ проект 35
3.12. Експортиране на резултатите от анализа 36
3.13. Сравняване на анализи от различни проекти. Анотиране на резултатите. 38
3.14. Вътрешни ограничения на софтуера 39
3.15. Системни изисквания 39
4. Пример за анализ на реални данни 40
4.1. Дизайн на биологичния експеримент и данни. 40
4.2. Проекти 43

4.3. Сравнения между отделните проекти. Резултати от анализа. 45
4.4. Сравнение между резултатите от анализа и резултати, получени от анализ на данните с друг софтуер за анализ на генна експресия 46
Използвана литература 51
Приложения 54
Приложение 1 – Флагове и контроли в CodeLink™ кДНК матрици 54
Приложение 2 – Анализ на реални данни от реален биологичен експеримент – детайли. 55
Приложение 3 – Инструкции за инсталиране 61
Приложение 4 – Разпечатки на екрани 62
Приложение 5 – Терминологичен речник 65
Приложение 6 – Експертно мнение на потенциални потребители 66
Приложение 7 – Аудио / видео презентация на софтуера и възможностите му на български език 67
Приложение 8 – Аудио / видео презентация на софтуера и възможностите му на английски език 67

Структура на дипломната работа

Глава 1 - Въведение – има за цел да предостави минимално необходимите познания и идеи за разбиране същността на разработката, предмет на настоящата дипломна работа. Въведението изяснява основни биологични понятия, дава идея за технологията на кДНК матриците и използването й, мотивира разработката и изяснява нейната цел и задачи.

Глава 2 – Биологична основа. ДНК. Гени. Генна експресия. Технология на кДНК матриците – изяснява биологичната основа, технологията на кДНК матриците, възможните грешки от технологията и стандартизацията на данните.

Глава 3 – Информационна система за анализ на генна експресия "zRMicroArray – Microarray Analysis tool" – детайлно описва разработения софтуер, неговите възможности, ограничения и функционалност, използваните програмни средства и статистически апарат и изяснява възможни насоки за развитие на разработката.

Глава 4 – Пример за анализ на реални данни – използване на софтуера за анализ на реални данни от реален биологичен експеримент. Сравнение на резултатите с резултати от друг софтуер за анализ на генна експресия.

Глава 5 – Заключение

1.Въведение

1.1. Генна експресия

Генна експресия е процес, при който унаследяемата информация от ДНК се трансформира във функционален продукт, който може да бъде:

- Матрична рибонуклеинова киселина (мРНК)
- Протеин, получен на базата на мРНК
- Транспортна РНК (тРНК)
- Рибозомна РНК (рРНК)
- Микро РНК (миРНК)

Количеството конкретни органични молекули, функционални продукти на ДНК ще наричаме ниво на генна експресия.

1.2. кДНК матрица (cDNA microarray)

Нивото на експресия на всеки един ген в генома на даден организъм в даден момент може да бъде отчетено посредством кДНК матрици (чипове).

1.3. Анализ на генна експресия

Различните организми от един и същи вид често се характеризират с различно ниво на експресия на гените, в резултат на въздействието на различни фактори на средата, което може да бъде отчетено посредством кДНК матрици. Получените данни трябва да бъдат обработени и статистически анализирани, за да бъдат открити диференциално

експресираните гени – тези, които имат статистически значимо различие в нивата на експресия при подлагане на организма на конкретни, различни експериментални фактори.

1.4. Мотивация

Диференциално експресираните гени, открити на база на статистически анализ на генна експресия, могат да бъдат използвани за:

- По-добро разбиране на биологичния механизъм на действие на различни заболявания на молекулярно ниво
- По-точно и бързо диагностициране на такива заболявания
- Откриване и проверка на различни диети и условия, които могат да повлияят позитивно или негативно по отношение на избраните критерии

Учените, най-вече молекулярни биолози, занимаващи се с експерименти, свързани с кДНК матрици, имат нужда от бърз, надежден и лесен за използване софтуер, посредством който да обработят и анализират данните, получени в резултат на дългия и скъпоструващ експеримент, който са провели.

1.5. Цел и задачи на дипломната работа

Цел на дипломната работа е проектиране и реализация на софтуерен продукт за статистически анализ на влиянието на различни фактори върху генната експресия на база на данни от кДНК матрици.

Основни задачи, произтичащи от целта са:

- Импортиране на данни от кДНК матрици
- Групиране на данните спрямо експерименталните условия
- Нормализация и филтриране на данните
- Статистически анализ на данните с цел откриване на диференциално екпресирани гени
- Експортиране на резултатите от анализа
- Създаване на подходящ, интуитивен графичен интерфейс
- Сравняване на резултати от анализи на близки експерименти

2. Биологична основа. ДНК. Гени. Генна експресия. Ниво на генната експресия. кДНК матрици. Стандартизация.

ДНК е дълга, най-често двуверижна органична молекула, съставена от голям брой линейно свързани нуклеотиди, всеки от които се състои от захар, фосфат и ароматно-въглеводородна база. Тъй, като базите са четири, то и нуклеотидите, изграждащи ДНК са четири:

- Аденин (А)
- Тимин (Т)
- Гуанин (Г)
- Цитозин (Ц)

Двете ДНК вериги се свързват една с друга на комплементарен принцип. По правило Аденин застава срещу Тимин, а Цитозин срещу Гуанин и обратно, което се определя от факта, че ароматно-въглеводородната база, съставяща конкретен нуклеотид, може да образува водородни връзки само с една от останалите три бази, съставящи другите нуклеотиди – А с Т и Г с Ц. На този принцип всяка от веригите може да бъде изградена на база на другата верига като матрица.
ДНК е носител на генетични инструкции за функционирането и развитието на всички известни живи организми и на някои вируси [11].
На база на матрицата ДНК се синтезират следните функционални генетични продукти:

- Матрична рибонуклеинова киселина (мРНК)
- Протеини, получени на базата на мРНК

- Транспортна РНК (тРНК)
- Рибозомна РНК (рРНК)
- Микро РНК (миРНК)

Кодиращата част от ДНК, носеща информацията за синтеза на конкретен функционален продукт, се нарича ген. Повечето гени представляват къси участъци от ДНК, които кодират информацията за синтеза на белтък. В гена има и регулаторни области – промотори (promoters) и енхансъри (enhancers), които не са кодиращи – те, в комбинация с геномни сигнали, определят кога и в какви количества да се синтезира съответният белтък или друг функционален продукт на ДНК.
Във всяка клетка има сложна регулаторна мрежа. Геномните сигнали, които участват в нея са [12]:

- ДНК
- мРНК
- миРНК
- Белтъци

Получените на база на ДНК конкретни функционални продукти ще наричаме генна експресия, а количеството на конкретна органична молекула - ниво на генната експресия.
Има различни технологии, които измерват нивото на генна експресия в конкретен момент.

Принцип на действие на технологията на кДНК матриците [12], [13], [14] и [15]:

1) Върху твърда повърхност, разделена на малки участъци – петна (spots), се прикрепят здраво къси, едноверижни ДНК последователности (олигонуклеотиди), които са

избрани така, че да са строго характерни[1] за конкретен, известен ген. Тази твърда повърхност с прикрепени към нея ДНК сегменти се нарича cDNA microarray (кДНК матрица). Една кДНК матрица може да съдържа десетки хиляди петна, към всяко от които са прикрепени голямо количество еднакви молекули.

2) След експериментално третиране или условия, примерно диета с рибено масло, фибри и/или раково заболяване, от тъкан се изолират всички молекули мРНК (mRNA transcriptomes).

3) Посредством ензимът обратна транскриптаза (РНК зависима ДНК полимераза) молекулите мРНК се трансформират на комплементарен принцип в кДНК транскрипти (Complementary DNA/cDNA reverse transcripts).

4) кДНК транскриптите се бележат с флуоресцентни багрила.

5) кДНК матрицата се залива с хомогенен течен разтвор на белязаните кДНК транскрипти, които се прикрепят (хибридизират) към съответните комплементарни ДНК последователности (фрагменти) върху матрицата, образувайки двуверижна ДНК чрез водородни връзки между двете вериги по принципа на комплементарността.

6) Матрицата се измива с химични агенти, за да се отстранят нехибридизиралите кДНК транскрипти.

[1] Последователностите са избрани така, че с тях да хибридизират единствено конкретно избрани кДНК транскрипти.

7) кДНК матрицата се сканира, като се осветява с лазер, при което флуоресцентното багрило започва да свети и излъчените фотони се регистрират с камера през микроскоп.

8) Снимката се разчита от софтуер за обработка на образи и в резултат се получава числена стойност за всяко едно петно, отчитаща интензитета на светене на съответното петно, който е пропорционален на броя молекули, хибридизирали към това петно, както и някои други параметри, отнасящи се към формата и големината на сканираните петна.

Описаната технология позволява да се отчете количеството на експресираните мРНК молекули в конкретен момент за всеки един ген, представен в кДНК матрицата. Понастоящем съществуват кДНК матрици, покриващи целия геном на някои модели организми, включително и такива, покриващи човешкия геном.

На всеки един от описаните етапи са възможни грешки от различно естество [18]:

- Непостоянства в повърхността на матрицата
- Допустимо отклонение на роботите (spotting pens), нанасящи петната
- Различно начално количество мРНК
- Различна наситеност на фосфоресциращото багрило
- Локални различия в условията на хибридизация вътре в кДНК матриците и между тях
- Различна калибровка на осветяващия лазер

- Различна чувствителност на камерата към дължината на вълната на излъчване на различните флуоресцентни бои
- Различна реализация на софтуера за обработка на образи и статистическа обработка
- Други

Поради тези грешки трябва да разглеждаме сигналите като случайни величини.
Нормализацията и трансформацията[1] на така получените данни се стреми да намали шума, получен в резултат на грешките и да направи данните подходящи за статистически анализ.

Поради:

- големия брой случайни величини в един кДНК чип,
- възможните грешки,
- скъпоструващата технология,
- необходимостта от повтаряне и проверка на експерименти, свързани с експресия на гени и
- необходимостта от сравняване на получените данни и резултати с други, получени от подобни експерименти

са необходими стандарти, като MIAME (Minimum Information About a Microarray Experiment), който [16], [17]:

- предоставя минималното количество информация, необходимо за повтаряне и проверка на експерименти, свързани с експресия на гени;
- не е точна спецификация, а дава основни насоки;
- е предпоставка за създаване на бази от данни за кДНК матрици и специализиран софтуер за обработката им.

[1] Вж. 3.9 - Трансформации

3. Информационна система за анализ на генна експресия “zRMicroArray – Microarray Analysis tool”

Повечето разработки в областта са web базирани и/или не разполагат с лесен за използване графичен интерфейс. Импортирането на данни обикновено става с много стъпки и файловете се импортират един по един. Обработките не са паралелно разпределени, което забавя многократно анализа при използване на съвременни, многоядрени процесори или многопроцесорни машини. Съществуват различни пакети, които предлагат различни функционалности и дават възможност за анализ на данни от кДНК матрици от различни производители. Повечето от тях не разполагат с графичен интерфейс и са подходящи само за потребители със значителни познания в областта на информационните технологии, програмирането и статистиката. Учените, най-вече молекулярни биолози, занимаващи се с експерименти, свързани с кДНК матрици, обикновено нямат такива познания и нямат възможността да обработят и анализират данните сами, което ги прави зависими от помощта на специалисти в областта на информационните технологии, програмирането и статистиката.

Настоящата разработка цели реализация на софтуерен продукт за статистически анализ на влиянието на различни експериментални фактори върху генната експресия, на база на данни от кДНК матрици, който да е бърз и лесен за използване на потребителско ниво.

3.1. Използвани програмни средства и дизайн на софтуера

За изграждане на информационната система са използвани следните програмни средства:

- Microsoft Visual FoxPro 9 sp 2
  - Графичен интерфейс
  - Бази от данни
  - Обработки

VFP (Microsoft Visual FoxPro [27] и [28]) е обектно ориентиран език за програмиране, ориентиран най-вече към релационни бази от данни и е подходящ за разработване на информационни системи в среда на Windows. Понастоящем правата за разпространение на VFP се държат от Microsoft. Разработването на информационна система с VFP изисква лицензирано копие на езика за програмиране. Разпространението на създадените с VFP програмни модули изисква наличието на runtime библиотеки (DLL), които не се заплащат допълнително. Информационните системи, разработени на VFP могат да бъдат използвани и под Linux, посредством пакета WineHQ [29], но това не се позволява от Visual FoxPro 9.0 EULA (end-user license agreement).

- R-project
  - Статистически анализ
  - Обработки
  - Графики

R-project е безплатен софтуер с отворен код за статистически изчисления и графики [30].

- R-(D)COM
    - Осъществява връзката между графичният интерфейс и R.

R-(D)COM е Microsoft distributed object interface за връзка с R и е безплатен за некомерсиални цели [31].

- WIN32API (стандартни Windows библиотеки)
    - Стартиране на паралелни процеси
    - Откриване на броя на процесорите и ядрата им

Основните функции на разработената информационна система са:

- Импортиране на данни от кДНК матрици
- Групиране на данните спрямо експерименталните условия
- Нормализация и филтриране на данните
- Статистически анализ на данните с цел откриване на диференциално екпресирани гени
- Експортиране на резултатите от анализа
- Сравняване на резултати от анализи на близки експерименти

Тази функционалност се изпълнява от следните програмни модули:

- zRMicroArray.exe – основен модул на информационната система. Изпълнява по-голямата част от гореописаната функционалност и предоставя на потребителя лесен за използване, интуитивен графичен интерфейс.

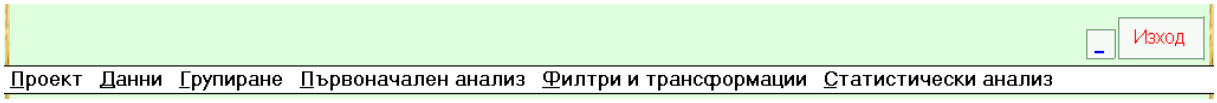
- zServer.exe – използва се от zRMicroArray за паралелно разпределени обработки

- zCodelinkConvert.exe – конвертира CodeLink™ .txt файлове във вид, подходящ за импортиране от системата

- zAnalysisCompare.exe – сравнява анализи от различни проекти

При стартиране на основния модул, програмата открива всички процесори и техните ядра и ги използва за паралелно разпределени обработки – т.е. първо задачите се разпределят на различен брой независими или зависими една от друга подзадачи, съобразени с броя на процесорите и техните ядра. Всяка подзадача се предава за изпълнение на отделен процесор или ядро. Задачата се смята за изпълнена след завършване на всяка от подзадачите и след обобщаване на резултатите от тях. Това значително ускорява анализа при използване на съвременни, многоядрени процесори или многопроцесорни машини. Паралелно се разпределят само обработки, отнемащи относително дълго време за връщане на резултат. Докато се извършва такава относително дълга обработка, системата не "заспива" и запазва голяма част от функционалността си[1], което уплътнява времето за извършване на анализа и позволява по-добра производителност на системата.

[1] Например: 1) докато се извършва импортиране на данни, потребителят може да разглежда различни статистики на вече импортираните данни, да въвежда фактори и стойности на тези фактори или даже да започне да групира данните. 2) Докато се извършва статистически анализ на конкретно филтрирани и трансформирани данни, потребителят може да генерира данни, филтрирани и трансформирани по друг начин.



фигура 1 - Главно меню на zRMicroArray

Основните екрани на информационната системата се стартират посредством главното меню (фигура 1) на системата и са следните:

- Информационен екран
- Проект
- Данни
- Групиране
- Предварителен анализ
- Филтри и трансформации
- Статистически анализ

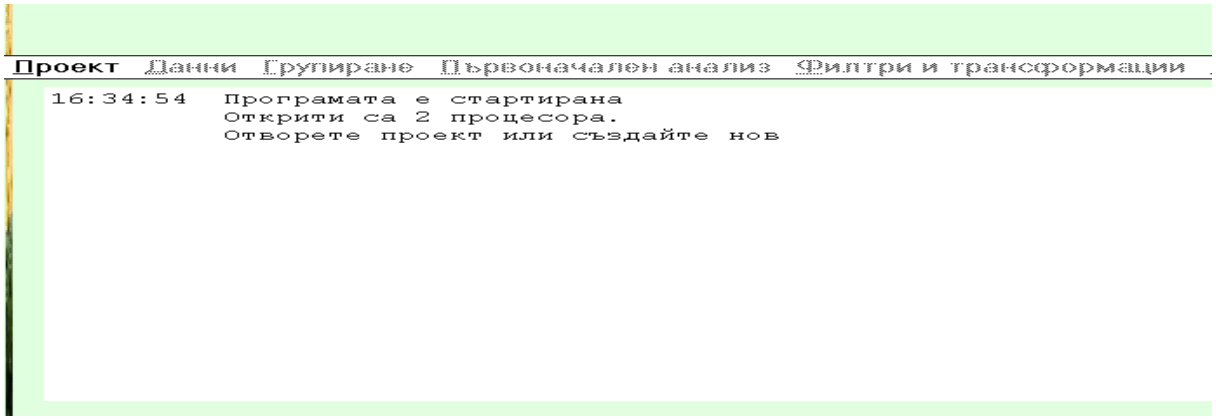
3.2. Превод на различни езици

Графичният интерфейс е преведен на български и английски език. Заложеният в системата речник позволява превод на до 254 езика. Понастоящем смяната на език изисква рестартиране на програмата.

3.3. Информационен екран

Информационният екран (фигура 2) има следните функции:

- Показва извършените от потребителя действия от момента на стартиране на програмата
- Информира потребителя за това какво се случва "зад екран" (in the background)
- Предлага подходящи по-нататъшни действия



фигура 2 - Информационен екран.

3.4. Меню “Проект”

Проектът е работната област за всеки експеримент. Меню „Проект” е единственото активно меню след стартиране на програмата и дава на потребителя следните възможности:

- Да създаде нов проект
- Да отвори вече съществуващ проект

Всички импортирани данни и направени обработки и анализи се записват автоматично в проекта в реално време.
При спиране на тока или при некоректно излизане от системата поради друга причина (рестартиране, бъг), системата се стреми:

- Да довърши започнатата работа или най-малкото
- Да запази текущото положение, без да наруши консистентността на базите от данни.

В една инстанция (instance) на информационната система, в един и същ момент, може да бъде отворен само един проект.

Системата обаче позволява стартирането на повече инстанции едновременно, давайки възможността за едновременна обработка на различни проекти.

3.5. Данни

3.5.1. Импортиране

Информационната система понастоящем може да импортира данни от CodeLink™ кДНК матрици в следните два формата:

- CodeLink™ TXT
- CodeLink™ XLS

Софтуерът на CodeLink™ няма определен файлов формат[1], а форматът се определя по време на експортиране. За да може системата да импортира данните директно, при експортиране от софтуера на CodeLink™ , трябва да бъдат избрани следните полета и то в същия ред:

- Idx
- Probe_name
- Probe_type
- Raw_intensity
- Normalized_intensity[1]
- Quality_flag

[1] Не се използва от софтуера и експортирането му не е задължително. Файлове без това поле трябва да бъдат предварително обработени от модула "zCodelinkConvert.exe".

Различните версии на CodeLink™ софтуера не само дават възможност за експортиране на различни полета, но и:

- Разбъркват последователността на полетата
- Експортират полетата с различни имена

Тези несъответствия във файловия формат водят до необходимостта от предварителен анализ и обработка на входните файлове. В информационната система могат да бъдат импортирани директно само изброените по-горе формати и то само експортирани по посочения начин. В останалите случаи е предвиден модулът "zCodelinkConvert.exe", който се грижи да намери съответните разбъркани полета и такива с променени имена и да ги прекодира в необходимия за импортиране формат. Модулът намира в данните и името на пробата (sample name), в което обикновено се съдържа и информация за експерименталните условия, на които е бил подложен опитният обект и генерира новия файл с име, съответстващо на името на пробата. По този начин се осигуряват „значещи" имена на файловете, което улеснява групирането на данните.
Понастоящем zCodelinkConvert.exe "разбира" CodeLink™ TXT файлове, експортирани по какъвто и да е начин от CodeLink™ софтуера.

3.5.2. Основни статистически показатели

Във всеки един момент след импортиране на данни, както и по време на импортирането, потребителят може да разгледа различни статистически показатели (фигура 3 и фигура 4) на вече импортираните файлове:

- Флагове[1], съдържащи се във файла и техният брой по конкретен флаг
- Контроли[1], съдържащи се във файла и броят на пробите по конкретна контрола

Stat	Count
C	15
CI	1
CL	15
G	26023
I	135
IS	4
L	9831
M	143
P	1
PI	1
S	58
L <= 0	254
DISCOVERY	34967
FIDUCIAL	640
NEGATIVE	320
POSITIVE	300
DISCOVERY <= 0	242
NEGATIVE <= 0	12
DISCOVERY = M	85
NEGATIVE = M	50
POSITIVE = M	8
Total	36227

фигура 3 - Статистики на данните. Флагове, контроли и техният брой.

- Некоректни данни по флагове и контроли - отрицателна генна експресия и контроли с липсващи стойности
- Общ брой на пробите
- Минимална стойност
- Максимална стойност
- Средна стойност
- Медиана
- Дисперсия (variance)

Min	Median	Mean	Max	Variance
0.0303	120.3584	832.0513	41157.8086	6157689.10

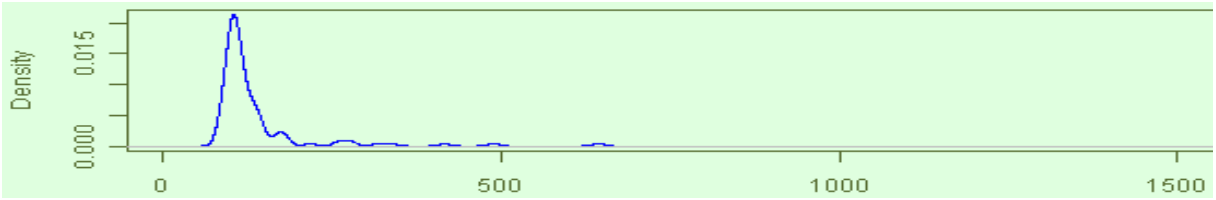
фигура 4 – Статистики на данните

3.5.3. Разпределение на данните и филтриране на ниво файл

Системата дава възможност за филтриране на данните на ниво файл и визуализация на разпределението на данните (фигура 5) в избрания филтър. Филтрирането на ниво файл се извършва в реално време, не е в резултат на предварителна обработка, визуализира се само на екран и има за цел запознаване на потребителя с конкретните данни, предмет на

[1] Вж. Приложение 1 - Флагове и контроли в CodeLink™ кДНК матрици

настоящия биологичен експеримент[1]. Филтрирането и визуализацията на разпределението е достъпно по всички статистики от фигура 3, а статистиките от фигура 4 се преизчисляват в реално време за избрания филтър.



фигура 5 - Разпределение на данните

3.5.4. Насоки за развитие

Модулът "zCodelinkConvert.exe" може да бъде развит така, че да конвертира повече файлови формати от повече платформи кДНК матрици във вид, подходящ за импортиране от основния модул на информационната система. Това ще разшири възможностите му за използване за анализ на повече видове данни.

Модулът може да бъде доразвит и така, че да конвертира данни, получени на база на новата deep sequencing технология, която директно брои експресираните молекули [24], [25] и [26].

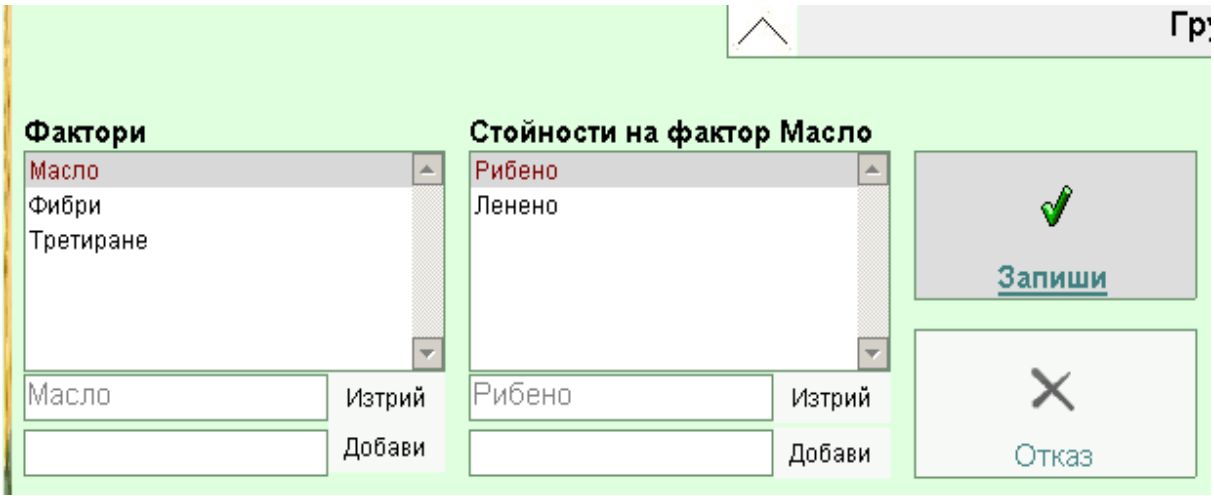
3.6. Групиране на данните от експеримента

Опитните биологични единици са подложени на различни експериментални условия. Тези разлики до голяма степен обуславят диференцираната генна експресия между

[1] Например потребителят може да разгледа данните за конкретна контрола или за конкретен флаг.

организмите. Информационната система предоставя възможност за описание на тези фактори на средата в две нива (фигура 6):

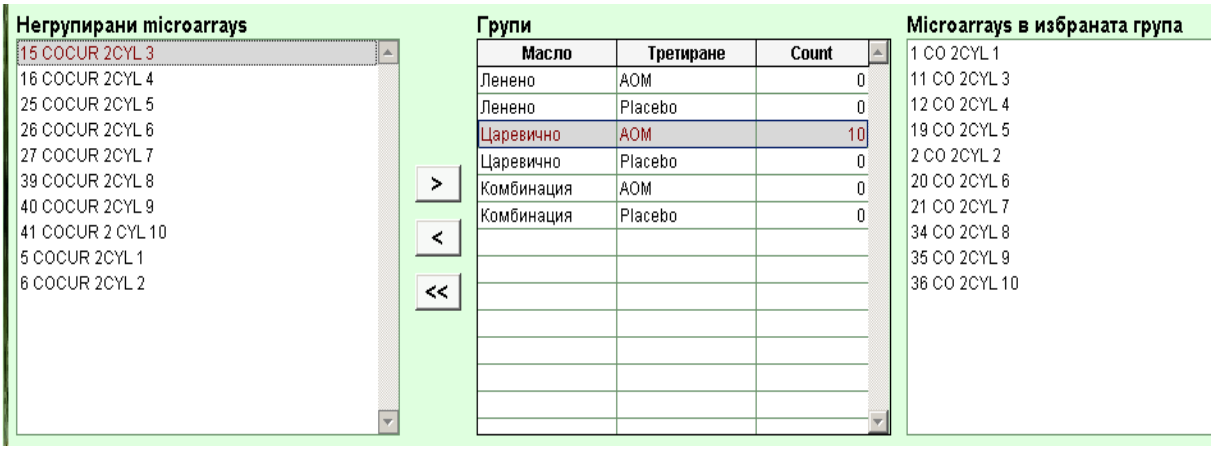
- Фактори
- Стойности на фактор



фигура 6 - Групиране. Фактори и нива на фактори.

Комбинацията от фактори и стойности на фактори описват различните групи, между които конкретният експеримент цели откриване на разликите в генната експресия при конкретните, различни фактори на средата.

Минимално допустимата комбинация е един фактор с две нива, при която групите са две. Обикновено обаче експериментът цели откриване на реакцията на гените, които представляват интерес на база на факторите поотделно, както и на комбинирания ефект на два и повече фактора (фигура 7).



фигура 7 - Групиране в съответствие с експерименталните условия.

3.7. Предварителен анализ

Предварителният анализ е съществена част от всяко статистическо изследване. Основна цел на предварителния анализ е запознаване с конкретните данни, предмет на настоящия анализ:

- идентифициране на данни, които не съответстват на останалите (outliers)
- преглед на разпределението на данните
- преглед на основни статистически показатели
- липсващи стойности и др.

3.7.1. Основни статистически показатели

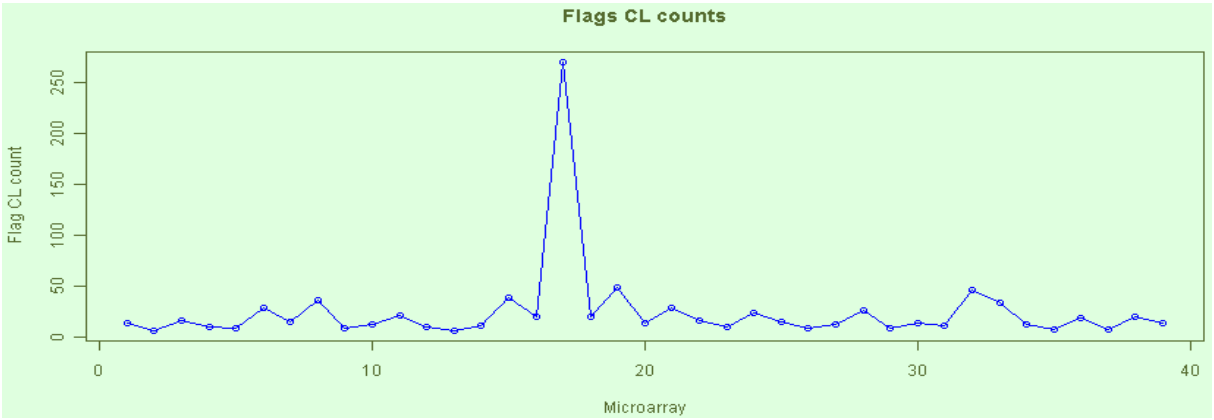
Информационната система zRMicroarray изчислява следните статистически показатели на данните и ги предоставя на потребителя таблично (фигура 8) и графично (фигура 9):

- Медиани
- Дисперсии
- Средни стойности

- Флагове – флагове, представени в отделните матрици и техният брой по съответен флаг

Microarray / Flags	Median	Mean	Variance	C	CI	CIS	CL	CS	G	I	IS	L	M	P	PC	PCI	PCL	PCLS	PCS	PI	PL	S
1 CO 2CYL 1	86.5609	665.5449	4724028	16	0	0	14	0	23542	136	4	12299	160	0	0	0	0	0	0	1	0	55
11 CO 2CYL 3	117.3700	657.4208	3120633	12	0	0	7	0	25840	97	0	10136	117	1	0	0	0	0	0	0	0	17
12 CO 2CYL 4	126.9048	909.3407	6624180	21	0	0	16	0	26247	106	6	9635	116	1	0	0	1	1	0	0	0	77
19 CO 2CYL 5	158.8649	959.7313	6919756	11	0	0	10	2	27456	93	5	8427	140	4	0	0	1	0	0	1	0	77

фигура 8 - Общи статистики. Таблично представяне.



фигура 9 - Общи статистики. Графично представяне.

3.7.2. Тест за нормално разпределение на гените в различните групи. Генериране на липсващи стойности.

Освен изчисляване на основните статистически показатели, в процеса на предварителен анализ, всеки ген се тества за нормално разпределение в принадлежащата му група. Използваният тест за нормално разпределение е Shapiro-Wilk normality test [5], [6] и [7].

За всеки ген, който според теста е нормално разпределен, се генерират липсващите стойности, ако такива съществуват. Генерира се случайна стойност от това нормално разпределение, която с по-голяма вероятност е близка до

медианата и средната стойност на разпределението и с по-малка вероятност е отдалечена от тях, но съобразено с дисперсията. За генериране на стойност се използва стандартна функция на R.

Решението дали така генерираните стойности да бъдат използвани или не за по-нататъшен статистически анализ, се взема от потребителя в екрана „Филтри и трансформации".
Тестът за нормално разпределение (Shapiro-Wilk) определя и изискването за минимум три кДНК матрици в една група.

3.7.3. Насоки за развитие

Подходящо е реализиране и на друг тест за нормално разпределение и по-конкретно D'Agostino-Pearson omnibus test [7]. Основната причина за това е, че Shapiro-Wilk тестът не работи добре, когато в данните има еднакви стойности. Не е характерно такива еднакви стойности да се появят в данни от кДНК матрици. Въпреки това изискването ограничава възможните методи за генериране на липсващи стойности[1].

3.8. Филтри

В екрана за филтри потребителят има възможност да избере, в зависимост от флаговете[2], кои гени да бъдат използвани в статистическия анализ.

[1] Вж. 3.7.2 - Тест за нормално разпределение на гените в различните групи. Генериране на липсващи стойности.
[2] Вж. Приложение 1 - Флагове и контроли в CodeLink™ кДНК матрици

Пример за филтриране по флагове и генериране на липсващи стойности:

Искаме да филтрираме данните, така, че да оставим само гените с флагове "G" и "M", като за "M" флагове (липсващи стойности) генерираме стойност по описаният в точка 3.7.2 начин:

ID на гени	Група 1	Група 1	Група 1	Група 2	Група 2	Група 2	Група 2
1	G	G	G	G	G	G	G
2	G	G	M (генерирана)	G	G	G	M (генерирана)
3	G	G	M (генерирана)	G	G	G	M
4	G	G	G	G	G	G	M (генерирана)
5	G	G	PSL	G	G	G	M (генерирана)

Легенда:

- **ID** на гена в **червено** - генът се **премахва** след филтриране.
- **ID** на гена в **зелено** - генът **остава** след филтриране.
- **Флагове** за цялата **група** в **червено** - стойностите в групата не са нормално разпределени.
- **Флагове** за цялата **група** в **зелено** - стойностите в групата са нормално разпределени.
- **Флагове** за цялата **група** в **черно** - в групата не присъстват липсващи стойности. Не се прави проверка за нормално разпределение.
- **Флаг** в **червено** – флагът не присъства в настоящия филтър.

Ген 1 – в данните присъстват **само "G"** флагове; **остава** след филтриране

Ген 2 – **група 1** е нормално разпределена и се генерират липсващи стойности; **група 2** е нормално разпределена и се генерират липсващи стойности; генът **остава** след филтриране.

Ген 3 – **група 1** е нормално разпределена и се генерират липсващи стойности; **група 2** не е нормално разпределена и не се генерират липсващи стойности; генът се **премахва** след филтриране.

Ген 4 – в **група 1** не присъстват липсващи стойности и не се прави проверка за нормално разпределение; **група 2** е нормално разпределена и се генерират липсващи стойности; генът **остава** след филтриране.

Ген 5 – в **група 1** присъства **флаг**, различен от посочените в конкретният филтър; **група 2** е нормално разпределена и се генерират липсващи стойности; генът се **премахва** след филтриране.

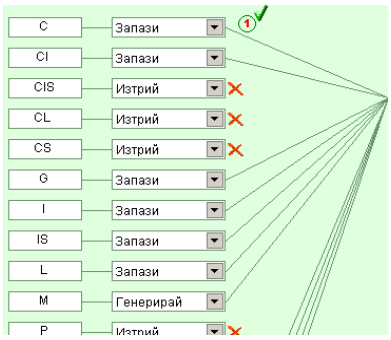
3.8.1. Контроли

Всички контроли се премахват от данните. Оставят се само гените, които представляват интерес[1] (Discovery)

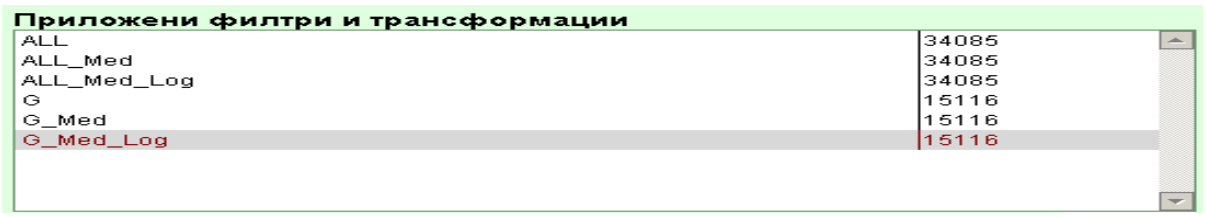
[1] Вж. Приложение 1 - Флагове и контроли в CodeLink™ кДНК матрици

3.8.2. Флагове

Информационната система открива всички флагове, съдържащи се в данните и дава възможност на потребителя да избере кои флагове да бъдат използвани за статистическия анализ (фигура 10). Потребителят може да генерира неопределен брой филтри (фигура 11) и да анализира получените филтрирани данни поотделно.



фигура 10 – Филтри. Избор на флагове.



фигура 11 - Приложени филтри и трансформации.

3.8.3. Насоки за развитие

Възможни са предварителни обработки на гените на база на флаговете, присъстващи в съответната група, с цел получаване на оценка за това до каква степен стойностите за съответния ген в съответната група са добре измерени и дали генът е подходящ за последващ статистически анализ. При присъствие на outliers (несъответстващи гени), дължащи се на лошо измерена стойност е възможно третиране на стойността като липсваща, или следва да се извърши друга предварителна обработка.

3.9. Трансформации

Съществуват три основни причини за трансформация на данните: постигането на нормално разпределение на данните, отделянето на дисперсията от средната стойност и редуцирането на различни не-адитивни взаимодействия. Разбира се изборът на „скала" на трансформация е емпиричен за постигане на определени цели – но самата трансформация е важен и често използван статистически инструмент, чиято основна цел е да се опрости максимално анализът на данните, до голяма степен тяхната интерпретация и отчитането на важни изводи.

В глава 2 – „Биологична основа" бяха разгледани възможни проблеми поради несъвършенства в технологията на кДНК матриците и човешки грешки. Нормализацията (трансформация) има за цел да филтрира шума от данните и да ги приближи до нормално разпределение, без да ги „изкривява", за да не се изгубят ефектите, които търсим [18], [19].

3.9.1. Реализирани трансформации

В системата са реализирани следните трансформации:

- Медианна трансформация – премахва различията в общото ниво на експресия между различните кДНК матрици.

- Логаритмична трансформация – приближава данните до нормално разпределение

3.9.2. Насоки за развитие

Реализиране на други трансформации [18], подходящи за данни от кДНК матрици:

- Lowess трансформация
- Нормализация по контроли и поддържащи гени (housekeeping genes)
- Други трансформации

3.10. Статистически анализ

Целта на използвания статистически подход е откриване на тези гени, които имат значимо различие в нивата на експресия при подлагане на организма на конкретни, различни експериментални условия.

Основните етапи на статистическия анализ са:

- Тест за нормално разпределение
- Дисперсионен анализ (Analysis of variance - ANOVA)
- Погрешност при идентифициране (False discovery rate - FDR)

3.10.1. Тест за нормално разпределение на гените в различните групи. Възможни комбинации от групи.

Всеки един ген, преминал заложените филтри, първо се тества за нормално разпределение в съответната му група. Това се прави поради изискването групите да са нормално

разпределени, за да бъдат тествани за различия, посредством дисперсионен анализ. Използваният тест за нормално разпределение е Shapiro-Wilk normality test [5], [6] и [7].
В екрана „Групиране" е определена принадлежността на всяка кДНК матрица към конкретна група. Групирането е направено на база на всички експериментални условия, на които е бил подложен организмът. Повечето гени се експресират под влиянието на един или няколко от въздействащите фактори, но не и в резултат на комбинация от тях. Това налага прегрупиране по всички възможни комбинации от фактори и тестването на всеки един вариант поотделно за различия в нивата на експресия на конкретен ген.

Пример:

Имаме три фактора със съответния им брой нива:

- Фактор1 – 2 нива
- Фактор2 – 2 нива
- фактор3 – 3 нива

Такава комбинация от фактори и техните нива обуславят дванадесет групи (2x2x3 = 12). Това е максималният брой групи, в които може да се разпределят данните от експеримента. Такова разпределение ще използваме само, когато търсим комбинирания ефект и на трите фактора заедно.
Възможните комбинации от фактори са:

- Фактор1 * Фактор2 * Фактор3 – 2x2x3 = 12 групи
- Фактор1 * Фактор2 – 2x2 = 4 групи
- Фактор1 * Фактор3 – 2x3 = 6 групи
- Фактор2 * Фактор3 – 2x3 = 6 групи
- Фактор1 - 2 групи
- Фактор2 - 2 групи
- Фактор3 – 3 групи

При три фактора са налице седем възможни комбинации за прегрупиране, всяка от които с определен, принадлежащ им брой групи. В конкретния случай ще трябва да тестваме всеки един ген за нормално разпределение в общо 12 + 4 + 6 + 6 + 2 + 2 + 3 = 35 различни групи.

3.10.2. Дисперсионен анализ (ANOVA)

Дисперсионният анализ е метод за търсене на статистически значими различия между две или повече групи нормално разпределени случайни величини. В конкретния случай ANOVA отговаря на въпроса за кои от гените имаме статистическо доказателство за разлики в нивото на експресия, поради подлагането на организма на въздействие на различни, конкретни експериментални условия. Различията между групите се обуславят от ефекта на факторите поотделно, както и на комбинирания им ефект.

В дадения в точка 3.10.1 пример разглеждаме седем различни комбинации от фактори, за да определим кои гени са диференциално експресирани възоснова на единични (базови) фактори или техни комбинации.

В резултат от ANOVA получаваме стойности на доверителния интервал (p-values[1]) за всеки един ген. При гранично ниво на p-value < 0.05 има 5% вероятност генът да не е диференциално експресиран.

В информационната система е използвана стандартната реализация на ANOVA в R [8] и [9].

[1] P value е вероятност със стойност между нула и едно, която отговаря на следния въпрос: Ако популацията наистина има една и съща средна стойност между групите, то каква е вероятността случайни величини да доведат до разлики между средните толкова голяма или по-голяма от наблюдаваната? [23]

3.10.3. Погрешност при идентифициране (False Discovery rate)

При анализ на много гени, 5%-ната вероятност генът да не е диференциално експресиран се мултиплицира. Ако открием 1000 диференциално експресирани гена, може да се очаква 50 от тях да са открити погрешно (false positives). Резултатът от FDR са стойности на доверителния интервал (q-values) на получените от дисперсионния анализ стойности за p-values. Информационната система използва реализирания FDR във fdrtool [10].

3.10.4. Насоки за развитие

- Реализиране на други тестове за нормално разпределение[1]
- Понастоящем системата търси диференциално експресирани гени, но няма механизъм за избор на контролни групи, спрямо които да анализира дали генът е с повишена (up-regulated) или намелена (down-regulated) експресия в отговор на експерименталните условия.

3.11. Сравняване на анализи от един и същ проект

Основният модул на информационната системата дава възможност на потребителя да сравнява резултатите от статистически анализи на различно филтрирани и трансформирани данни. Като резултат, потребителят

[1] Вж. 3.7.3 – Насоки за развитие

получава списък от общи или сумарни гени между различните анализи.

3.12. Експортиране на резултатите от анализа

Информационната система дава възможност за експортиране на резултатите от статистическите анализи, както и на резултатите от сравнения между анализите в следните файлови формати:

- Comma Separated Values (.csv)
- Microsoft Excel workbook file (.xls)
- Visual Fox Pro relational table (.dbf)

Експортирането се извършва в три колони:

- Колона 1 - Цифров идентификатор
- Колона 2 - Име на пробата от кДНК матрицата
- Колона 3
    - Фактор / фактори, по които генът е диференциално експресиран и
    - статистическа значимост спрямо таблицата:

Граници на p/q-value	Означение[1]
(0.05, 0.1]	'.'
(0.01, 0.05]	'..'
(0.001, 0.01]	'*'
(0.00001, 0.001]	'**'
[0, 0.00001]	'***'

[1] Означенията за статистическа значимост са заимствани от R с лека модификация ('.' вместо ' ' и '..' вместо '.') и са се наложили в практиката.

- Примери за експортиране за един ген в CSV формат:

639,"GE42467","(CURCUMINE[1] ' *[2] ')"
15566,"GE36273","(OIL ' ** ') (CUR ' .. ')"
4550,"GE1526068","(OIL &[3] CURCUMINE ' .. ') !!![4] (OIL ' .. ')"

- Пример за експортиране на сравнение между два анализа за един ген в CSV формат:

551,"GE122989","(CURCUMINE ' * ') |[5] (CURCUMINE ' ** ')"

Системата дава възможност и за отделно експортиране на откритите диференциално експресирани гени по конкретен фактор или комбинация от фактори.

[1] Фактор/фактори, по който генът е диференциално експресиран
[2] Статистическа значимост
[3] Изразява комбиниран ефект на фактори
[4] Внимание! Не интерпретирайте ефекта на базов фактор, при наличие на взаимодействие между факторите! [20], [21]
[5] Разделяща линия между два различни анализа при сравнение

3.13. Сравняване на анализи от различни проекти. Анотиране на резултатите.

В информационната система е предвиден модул „zAnalysisCompare”, който позволява сравняване на резултатите от анализи на различни експерименти. Модулът дава и възможност за анотиране на откритите диференциално експресирани гени спрямо различни биологични анотации (NCBI, LocusLinc, UniGene, ENSEMBL и други).
Понастоящем модулът е със завършена функционалност, но с непълно завършен графичен интерфейс.

На вход модулът приема:

- анотации в CSV формат с две колони, разделени с SPACE, TAB или comma:
  - Колона 1 – име на пробата
  - Колона 2 – биологична анотация

- Неограничен брой входни списъци с гени, експортирани от основния модул на системата.

Резултатът от сравнението е списък с биологично анотирани гени, получени на база на:

- Общи гени между анализите
- Сумарни гени от анализите
- Гени, присъстващи в част от входните списъци с гени, но неприсъстващи в друга част от тях.

3.14. Вътрешни ограничения на софтуера

В информационната система са заложени следните ограничения:

	максимум	минимум
Брой:		
• кДНК матрици	∞	
• проби в една кДНК матрица	99999	
• групи	9999	2
• кДНК матрици в една група	999	3
• фактори[1]	254	1
• нива на фактори	∞	2
Брой символи за:		
• име на фактор[1]	10	1
• име на ниво на фактор	10	1
• име на файл за импортиране	100	

3.15. Системни изисквания

- Операционна система:
  - Windows 98, Windows Me, Windows 2000 Service Pack 2 or later, Windows XP, Windows XP 64x, Windows Vista, Windows Vista 64x, Windows 7.

- RAM памет:
  - 90 MiB (mebibyte)
  - Допълнителни 22 MiB за всеки допълнителен процесор и всяко допълнително ядро на процесор.

[1]Ограничението е силно свързано с вътрешния дизайн на софтуера и трудно може да бъде променено в настоящата имплементация.

4.Пример за анализ на реални данни

Ще разгледаме анализ на реални данни от реален биологичен експеримент. Експериментът не е стандартен и към него не може да бъде приложен стандартен подход за анализ. По-детайлно описание на направения анализ е дадено в Приложение 2 - Анализ на реални данни от реален биологичен експеримент – детайли.

4.1. Дизайн на биологичния експеримент и данни.

Биологичният експеримент е проведен по следния начин:

От Jackson Laboratories (Bar Harbor, ME) са взети мъжки мишки (C57BL/6) на възраст между шест и осем седмици. Мишките са аклиматизирани една седмица при следните условия:

- 12 часов цикъл на светло и тъмно в контролирана от гледна точка на влажност и температура среда

- Стандартна диета на гранулирана храна

Опитните животни са разделени произволно на групи от по 15 екземпляра и са хранени с една от четири различни диети (2x2 дизайн), които се различават единствено по мастната киселина, със или без добавен куркумин (активно вещество, извлечено от подправката куркума).

Четирите диети, спрямо които опитните животни са разделени на групи са:

- Рибено масло (FO)
- FO+2% куркумин
- Царевично олио (CO)
- CO+2% куркумин

Една седмица след разделянето на опитните животни и подлагането им на четири различни диети, са взети проби от тъкан от дебелото черво от една от групите – тази на диетата с CO.
Мишките от всяка от диетите са пили от разтвор на DSS[1] (dextran sodium sulfate) в продължение на 5 дни, след което са пили чиста вода в продължение на 16 дни, продължавайки да се хранят по съответната диета. След това мишките са били подложени на още един, тридневен цикъл с по-ниска концентрация на DSS и на възстановителен период от 14 дни.
След възстановителния период мишките са били евтанизирани хуманно и дебелите им черва са били отстранени.

От тъкан от дебелото черво са изолирани всички мРНК транскрипи и следвайки процеса, описан в глава 2 – биологична основа, са получени данни от 45 кДНК матрици в 5 групи:

- COcon – Контролна група от 5 кДНК матрици, получени от мишки, хранени с диетата с CO. При предварителен анализ е открита 1 кДНК матрица, която е със значително по-високи средна стойност, медиана и дисперсия и със значително повече "S" и "C" флагове, сравнени с

[1] Причинява тумори на дебелото черво

останалите 44 кДНК матрици. Този проблем в данните ще бъде отстранен посредством прилагане на медианна трансформация. Не са открити диференциално експресирани гени между групите със и без данните за тази кДНК матрица. Поради тази причина и поради малкия брой кДНК матрици в групата, тези данни ще бъдат използвани за по-нататъшен анализ.

- CO_DSS – 10 кДНК матрици, получени от мишки, хранени с диетата с CO и третирани с DSS

- CO_CUR_DSS – 11 кДНК матрици, получени от мишки, хранени с диетата с CO с добавен куркумин и третирани с DSS. При първоначален анализ е открита 1 кДНК матрица, която е със значително по – малко "G" флагове[1]. Тези данни няма да бъдат използвани за по-нататъшен анализ.

- FO_DSS – 9 кДНК матрици, получени от мишки, хранени с диетата с FO и третирани с DSS

- FO_CUR_DSS – 10 кДНК матрици, получени от мишки, хранени с диетата с FO с добавен куркумин и третирани с DSS

В така получените данни (общо 5 групи) има 11 различни ефекта на заложените в експеримента условия - CO, FO, CUR, DSS, CO&[2]CUR, FO&CUR, CO&DSS, FO&DSS, CUR&DSS, CO&CUR&DSS, FO&CUR&DSS. Не всички от тях могат да бъдат изучени в този експериментален дизайн, поради липсата на три контролни групи (FO, FO_CUR, CO_CUR) .

[1] Вж. Приложение 1 - Флагове и контроли в CodeLink™ кДНК матрици
[2] & означава комбиниран ефект

4.2. Проекти

При експеримент, при който са заложени такива експериментални условия, в информационната система „zRMicroArray" би следвало да се създаде проект, да се въведат три фактора, всеки от които да има по две значения, данните да се разпределят в получените осем групи и да се направи статистически анализ посредством мултифакторен дисперсионен анализ.

Случаят, обаче, не е такъв и за анализиране на данните е необходимо създаването на следните шест проекта:

- COcon срещу CO_DSS – един фактор с две значения – общо 2 групи

- COcon срещу FO_DSS[1] – един фактор с две значения – общо 2 групи

- COcon срещу CO_CUR_DSS[1] – един фактор с две значения – общо 2 групи

- COcon срещу FO_CUR_DSS[1] – един фактор с две значения – общо 2 групи

- COcon срещу всички останали групи - един фактор с две значения – общо 2 групи – Можем да разглеждаме така получените диференциално експресирани гени като такива, върху които DSS оказва влияние, но FO, CO и CUR, нито някой от комбинираните им ефекти не променят

[1] В така получените диференциално експресирани гени остават ефекти, които са нежелани, но няма как да изчистим шума от тях.

това влияние. Поради малкия брой кДНК матрици, трябва да внимаваме с интерпретацията на този списък от гени, особено при наличие само на една контролна група.

- DSS третирани групи – два фактора, всеки от които с по две значения – общо 4 групи – не очакваме всичките диети да оказват влияние върху туморите на дебелото черво, нито това да се случва по един и същи начин. Можем да премахнем от горните списъци гените, които не са диференциално експресирани между DSS третираните групи.

4.3. Сравнения между отделните проекти. Резултати от анализа [32].

Гореполучените списъци с гени и различни комбинации между тях могат да отговорят на различни биологични въпроси.

Въпрос:
Кои са гените, върху които DSS има статистически значим ефект, а добавянето на куркумин премахва този ефект, регулирайки генната експресия по такъв начин, че да я връща в състоянието преди третиране с DSS?

Отговор (фигура 12):
Търсените гени са диференциално експресирани спрямо контролната група (означено със зелено на фигура 12) при третиране с DSS и добавяне към диетата на FO или CO, но не са диференциално експресирани спрямо контролната група при добавяне и на куркумин (означено с черно на фигура 12).

	COcon	CO_DSS	CO_CUR_DSS	FO_DSSl	FO_CUR_DSS	
1						Ген 1
2						Ген 2
3						Ген 3
4						Ген 4
5						Ген 5
6						Ген 6
7						Ген 7
8						Ген 8
9						Ген 9
10						Ген 10
11						Ген 11

фигура 12 - Гени, при които куркуминът неутрализира ефекта на DSS

Отговори на други биологични въпроси, на които може да отговори този биологичен експеримент, можете да намерите в Приложение 2 – Анализ на реални данни от реален биологичен експеримент – детайли.

4.4. Сравнение между резултатите от анализа и резултати, получени от анализ на данните с друг софтуер за анализ на генна експресия

В таблица 1, таблица 2 и таблица 3 са показани сравнения между резултатите, получени от статистически анализ на данните с информационната система "zRMicroArray – Microarray Analysis tool", при различни настройки и web – базирания софтуер GeneSifter [22].

	Брой диференциално експресирани гени		
Сравнение 1	**GeneSifter**	**zRMicroArray**	**Общи**
Настройки, филтри и трансформации	Всички флагове; медианна и логаритмична трансформация, гранична FDR q-value 0.05;	G и M флагове; медианна и логаритмична трансформация, гранична FDR q-value 0.05;	**45%**[1]
CO_CON срещу CO_DSS	526	160	119
CO_CON vs CO_DSS_CUR	1366	1387	648
CO_CON vs FO_DSS	609	418	231
CO_CON vs FO_DSS_CUR	1471	2635	807

таблица 1 - сравнение между zRMicroarray и GeneSifter

[1] Процент на съвпадение при сумиране по колони.

	Брой диференциално експресирани гени		
Сравнение 2	**GeneSifter**	**zRMicroArray**	**Общи**
Настройки, филтри и трансформации	Всички флагове; медианна и логаритмична трансформация, гранична FDR q-value 0.05;	Всички флагове; медианна и логаритмична трансформация, гранична FDR q-value 0.05;	**62%**
CO_CON срещу CO_DSS	526	218	192
CO_CON срещу CO_DSS_CUR	1366	1226	867
CO_CON срещу FO_DSS	609	285	220
CO_CON срещу FO_DSS_CUR	1471	2207	1152

таблица 2 - сравнение между zRMicroarray и GeneSifter

	Брой диференциално експресирани гени		
Сравнение 3	**GeneSifter**	**zRMicroArray**	**Общи**
Настройки, филтри и трансформации	Всички флагове; медианна и логаритмична трансформация, гранична FDR q-value 0.05;	Всички флагове; медианна и логаритмична трансформация; без използване на тест за нормално разпределение; гранична FDR q-value 0.05;	**66%**
CO_CON срещу CO_DSS	526	235	210
CO_CON срещу CO_DSS_CUR	1366	1286	942
CO_CON срещу FO_DSS	609	294	235
CO_CON срещу FO_DSS_CUR	1471	2279	1240

таблица 3 - сравнение между zRMicroarray и GeneSifter

Забелязваме различия между откритите гени от двата софтуера. Тези различия са в резултат на:

- Различна предварителна обработка на флаговете
- Използването или не на тест за нормално разпределение
- Различно претегляне на извадките
- Различия в използваният FDR[1] метод

Най-близки са резултатите, когато информационната система zRMicroArray е настроена така, че да не използва тест за нормално разпределение на гените по групи, което води до извода, че GeneSifter не използва тест за нормално разпределение, въпреки, че дисперсионният анализ изисква данните да са нормално разпределени.

[1] GeneSifter използва Benjamini-Hochberg false discovery rate

5. Заключение

Дипломната работа има за цел и изпълнява следните задачи:

- Импортиране на данни от кДНК матрици.

Информационната система импортира данни от CodeLink™ кДНК матрици. Допълнителен модул дава възможност за доразвиване на системата с оглед на импортиране на повече платформи и файлови формати.

- Групиране на данните спрямо експерименталните условия.

Реализираният модел за описание на експерименталните условия на две нива дава възможност за стандартно групиране на данните от различни биологични експерименти.

- Нормализация и филтриране на данните.

Разработката дава възможности за различни трансформации и филтрирания на данните.

- Статистически анализ на данните с цел откриване на диференциално екпресирани гени по данни от кДНК матрици.

Използван е подходящ статистически апарат, посредством който системата открива диференциално експресирани гени спрямо различни експериментални условия.

- Експортиране на резултатите от анализа.

Резултатите от статистическия анализ се експортират в удобен и информативен вид и в различни файлови формати.

- Създаване на подходящ, интуитивен графичен интерфейс.

Създаден е интуитивен графичен интерфейс, даващ на потребителя бърз и лесен начин за анализиране на данни от кДНК матрици.

- Сравняване на резултати от анализи на близки експерименти

Допълнителен модул дава възможност за сравняване на данни от близки експерименти. Модулът дава възможност и за анотиране на получените резултати спрямо различни биологични анотации.

Основни предимства на настоящата разработка пред другите разработки в същата област са:

- Паралелно разпределени обработки
- Удобен за използване, интуитивен графичен интерфейс

Информационната система е тествана върху различни реални данни и е съгласувана с потенциални потребители[1].

Получените резултатите са сравнени с резултати от друг софтуер и различията са обяснени и описани.

Поставената цел да бъде проектирана и реализирана информационна система за анализ на генна експресия на базата на данни от кДНК матрици е изпълнена.

[1] Вж. Приложение 6 – Експертно мнение на потенциални потребители

Използвана литература

[1] Codelink: an R package for analysis of GE healthcare gene expression bioarrays http://bioinformatics.oxfordjournals.org/cgi/content/full/23/9/1168

[2] Diego Diez, Codelink, April 21, 2009 http://www.bioconductor.org/packages/2.4/bioc/vignettes/codelink/inst/doc/codelink.pdf

[3] CodeLink™ Expression Analysis Software Version 5.0 http://www.appliedmicroarrays.com/pdf/CodeLink_Software_Version_5.pdf

[4] CodeLink™ Bioarray Image Threshold Flagging Procedure When Using the Agilent G2565BA Scanner http://www.appliedmicroarrays.com/application_notes/28903376aa.pdf

[5] Shapiro, S. S. and Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)", *Biometrika*, 52, 3 and 4, pages 591–611 http://sci2s.ugr.es/keel/pdf/algorithm/articulo/shapiro1965.pdf

[6] Shapiro-Wilk Normality Test in CRAN http://cran.us.r-project.org/doc/manuals/R-intro.html#Examining-the-distribution-of-a-set-of-data

[7] Normality tests – use with caution http://www.graphpad.com/library/BiostatsSpecial/article_197.htm

[8] Julian J. Faraway, Practical Regression and Anova using R, July 2002 http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf

[9] R Guide -- Analysis of Variance http://www.personality-project.org/r/r.anova.html

[10] Korbinian Strimmer, A unified approach to false discovery rate estimation, 2008
http://www.biomedcentral.com/1471-2105/9/303
[11] DNA – Wikipedia
http://en.wikipedia.org/wiki/DNA
[12] Иван Иванов, Advanced computational biology, lectures, Texas A&M University, Софийски университет „Св. Климент Охридски", юли 2008, юли 2009
[13] DNA microarray – wikipedia
http://en.wikipedia.org/wiki/DNA_microarray
[14] Gary Hardimen, Microarray platforms – comparisons and contrasts, Department of Medicine at UCSD, Biomedical Genomics laboratory, 2004
http://microarrays.ucsd.edu/biogem/pdf/Hardiman%20G.pdf
[15] Steven M. Carr, cDNA microarrays, 2008
http://www.mun.ca/biology/scarr/cDNA_microarray_Assay_of_Gene_Expression.html
[16] Minimum information about a microarray experiment (MIAME)—toward standards for microarray data
http://www.frontiers-in-genetics.org/documents/gen-services/miame_standards.pdf
[17] MIAME Home page
http://www.mged.org/Workgroups/MIAME/miame.html
[18] John Quackenbush, Microarray data normalization and transformation, The Institute for Genomic Research, Maryland, USA, 2002
http://www.nature.com/ng/journal/v32/n4s/abs/ng1032.html
[19] John Quackenbush, Computational analysis of microarray data, Macmillan Magazines, 2001
http://skop.genetics.wisc.edu/AudreyWeek6/Quackenbush_01.pdf
[20] Gerard E. Dallal, The Model For Two-Factor Analysis of Variance, Ph.D, 2006

http://www.jerrydallal.com/LHSP/anova2model.htm

[21] Klaus Hinkelmann, Evaluating And Interpreting Interactions, December 13, 2004
http://www.stat.org.vt.edu/dept/web-e/tech_reports/TechReport04-6.pdf

[22] Genesifter – official site
http://www.genesifter.net/

[23] What is a P value?
http://www.graphpad.com/articles/pvalue.htm

[24] Next generation sequencing
http://www.agencourt.com/services/nextgen/

[25] Wikipedia - DNA sequencing, High-throughput sequencing
http://en.wikipedia.org/wiki/DNA_sequencing#High-throughput_sequencing

[26] Wikipedia - RNA-Seq
http://en.wikipedia.org/wiki/RNA-Seq

[27] Microsoft Visual FoxPro Home
http://msdn.microsoft.com/en-us/vfoxpro/default.aspx

[28] Wikipedia – Visual FoxPro
http://en.wikipedia.org/wiki/Visual_FoxPro

[29] WineHQ Home
http://www.winehq.org/

[30] R-project Home
http://www.r-project.org/

[31] R-(D)COM / R-Excel
http://www.sciviews.org/_rgui/projects/RDcom.html

[32] Dietary effects of curcumin supplementation on collon cancer development using a DSS mouse model, J. Qian, J. Goldsby, Z. Zlatev, I. Ivanov, R. Chapkin, in preparation

Приложения

Приложение 1 – Флагове и контроли в CodeLink™ кДНК матрици [2],[3],[4]

- Контроли:
  - DISCOVERY – Проби от интерес
  - POSITIVE – Положителни контроли – за тях трябва да има добре измерен сигнал
  - NEGATIVE – Отрицателни контроли - за тях не трябва да има добре измерен сигнал и стойностите трябва да са ниски. Отчитат дали биологичната проба е замърсена.
  - FIDUCIAL - Използват се при сканиране на кДНК матрици (Grid alignment)
  - OTHER – други контроли и поддържащи гени

- Флагове:
  - G – Добре измерен сигнал
  - L - Слаб сигнал
  - I – Неправилна форма на "петното" (spot)
  - S – Наситен сигнал
  - M – Липсваща стойност
  - C – Замърсеност на кДНК матрицата
  - P – Замърсеност на петното (нов флаг от версия 5.0 на CodeLink™ Expression Analysis Software)
  - X – Петна, изключени от анализа от страна на потребителя

  Всяко петно може да съдържа комбинация от един, два или повече флага.

Приложение 2 – Анализ на реални данни от реален биологичен експеримент – детайли.

1. **Data & Groups**
   1.1. 45 microarrays
   1.2. 5 groups:

      1.2.1. CO_CON – 5 microarrays:

T00343454	Extremely high mean, median and variance compared to all other (44) MA's. Mach more S and C flagged genes. I wouldn't use it if there ware more microarrays in the control group. This effect will be removed by the median transformation. T-test between 5 microarrays control group and 4 microarrays (without T00343454) done with P-Value 0.05, and median transformed data. No differentially expressed genes found between this 2 "groups".

      1.2.2. CO_DSS – 10 microarrays

      1.2.3. FO_DSS – 9 microarrays

      1.2.4. CO_CUR_DSS – 11 microarrays

T00343402	Extremely low "G" flagged genes count. Microarray is totally excluded from the analysis

      1.2.5. FO_CUR_DSS – 10 microarrays

2. There are 12 effects in the data (all 5 groups) - CO, FO, CUR, DSS, CO*CUR, FO*CUR, CO*DSS, FO*DSS, CUR*DSS, CO*CUR*DSS, FO*CUR*DSS ('*' used for combined effect). They can't be studied all in experiments designs like this. There should be control groups for all diets, so in this experiment there should be 8 groups: FO, CO, FOCUR, COCUR, FODSS, CODSS, FODSSCUR and CODSSCUR. It's OK to compare CO (Control group) with CO_DSS, but when we compare CO with FO_DSS there is noise in the results and we can't rely on them.

3. **All data is median and logarithmic transformed; 0.05 cutoff Q-value used for FDR; Only normally distributed genes are passed to ANOVA; Missing values were generated for normally distributed genes.**

4. Multifactorial ANOVA effects - this could be tested only for DSS treated groups, because no microarray data presented for 3 of the groups, not affected by DSS.
   4.1. G&M(generated) flagged:
      4.1.1. **1662** genes – "all2wayGM.csv"
   4.2. All flagged:
      4.2.1. **1554** genes – "all2wayALL.csv"
   4.3. Aggregate between 4.1.1 and 4.2.1 – "2wayAggregate.csv" – **2360**
   4.4. Mutual between 4.1.1 and 4.2.1 – "2wayMutual.csv" – **856**
   4.5. **We don't expect all the diets to reduce the DSS effect neither the same way, so we expect that only differentially expressed genes on the different diets will be the difference. We'll use "2wayAggregate.csv" to filter CONvs files** (see below) **(only this genes stay)**

5. **DSS effect, which is not changed by CO, FO or CUR** supplementation - CO_CON (5 microarrays) vs CO_DSS, CO_CUR_DSS, FO_DSS, FO_CUR_DSS grouped together (39 microarrays):
   5.1. G&M(generated) flagged:
      5.1.1. **639** genes – "DSSGM.csv"
   5.2. All flagged:
      5.2.1. **924** genes – "DSSALL.csv"
   5.3. Aggregate between 5.1.3 and 5.2.3 – "DSSAggregate.csv" – **1104**
   5.4. Mutual between 5.1.3 and 5.2.3 – "DSSMutual.csv" - **459**
   5.5. **We may consider genes in "DSSAggregate.csv" and "DSSMutual.csv" as genes effected by DSS treatment, but not effected by CO, FO or CUR supplementation neither by some of their combined effects** (contained in the data)**. We'll use "DSSMutual.csv" to filter CONvs files** (see below) **(this genes are removed)**

6. CO_CON **vs CO_DSS**
   6.1. G&M(generated) flagged:
      6.1.1. **218** - "CONvsCODSS_GM.csv"
   6.2. All flagged:
      6.2.1. **160** - "CONvsCODSS_ALL.csv"
   6.3. Aggregate between 6.1.1 and 6.2.1 – "CONvsCOAggregate.csv" – **266**
   6.4. Mutual between 6.1.1 and 6.2.1 – "CONvsCOmutual.csv" – **112**
   6.5. Filtered "CONvsCODSS_GM.csv" -> 2way filtered (stay) -> DSS effect filtered (removed) -> **19** genes -> "CONvsCO_GM_Filtered.csv"
   6.6. Filtered "CONvsCODSS_ALL.csv" -> 2way filtered (stay) -> DSS effect filtered (removed) -> **23** genes -> "CONvsCO_ALL_Filtered.csv"

6.7. Filtered "CONvsCOAggregate.csv" -> 2way filtered (stay) -> DSS effect filtered (removed) -> **30** genes -> "CONvsCOaggFiltered.csv"
6.8. Filtered "CONvsCOmutual.csv" -> 2way filtered (stay) -> DSS effect filtered (removed) -> **12** genes -> "CONvsCOmutualFiltered.csv"

7. CO_CON **vs FO_DSS**
   7.1. G&M(generated) flagged:
      7.1.1.**418** - "CONvsFODSS_GM.csv"
   7.2. All flagged:
      7.2.1.**285** - "CONvsFODSS_ALL.csv"
   7.3. Aggregate between 7.1.1 and 7.2.1 – "CONvsFOAggregate.csv" – **518**
   7.4. Mutual between 7.1.1 and 7.2.1 – "CONvsFOmutual.csv" – **185**
   7.5. Filtered "CONvsFODSS_GM.csv" -> 2way filtered (stay) -> DSS effect filtered (removed) -> **35** genes -> "CONvsFO_GM_Filtered.csv"
   7.6. Filtered "CONvsFODSS_ALL.csv" -> 2way filtered (stay) -> DSS effect filtered (removed) -> **30** genes -> "CONvsFO_ALL_Filtered.csv"
   7.7. Filtered "CONvsFOAggregate.csv" -> 2way filtered (stay) -> DSS effect filtered (removed) -> **48** genes -> "CONvsFOaggFiltered.csv"
   7.8. Filtered "CONvsFOmutual.csv" -> 2way filtered (stay) -> DSS effect filtered (removed) -> **17** genes -> "CONvsFOmutualFiltered.csv"

8. CO_CON **vs COCUR_DSS**
   8.1. G&M(generated) flagged:
      8.1.1.**1387** - "CONvsCOCURDSS_GM.csv"
   8.2. All flagged:
      8.2.1.**1226** - "CONvsCOCURDSS_ALL.csv"
   8.3. Aggregate between 8.1.1 and 8.2.1 – "CONvsCOCURAggregate.csv" – **1835**
   8.4. Mutual between 8.1.1 and 8.2.1 – "CONvsCOCURmutual.csv" – **778**
   8.5. Filtered "CONvsCOCURDSS_GM.csv" -> 2way filtered (stay) -> DSS effect filtered (removed) -> **364** genes -> "CONvsCOCUR_GM_Filtered.csv"
   8.6. Filtered "CONvsCOCURDSS_ALL.csv" -> 2way filtered (stay) -> DSS effect filtered (removed) -> **298** genes -> "CONvsCOCUR_ALL_Filtered.csv"
   8.7. Filtered "CONvsCOCURAggregate.csv" -> 2way filtered (stay) -> DSS effect filtered (removed) -> **455** genes -> "CONvsCOCURaggFiltered.csv"

8.8. Filtered "CONvsCOCURmutual.csv" -> 2way filtered (stay) -> DSS effect filtered (removed) -> **207** genes -> "CONvsCOCURmutualFiltered.csv"

9. CO_CON **vs FOCUR_DSS**
   9.1. G&M(generated) flagged:
      9.1.1. **2635** - "CONvsFOCURDSS_GM.csv"
   9.2. All flagged:
      9.2.1. **2207** - "CONvsFOCURDSS_ALL.csv"
   9.3. Aggregate between 9.1.1 and 9.2.1 – "CONvsFOCURAggregate.csv" – **3297**
   9.4. Mutual between 9.1.1 and 9.2.1 – "CONvsFOCURmutual.csv" – **1545**
   9.5. Filtered "CONvsFOCURDSS_GM.csv" -> 2way filtered (stay) -> DSS effect filtered (removed) -> **698** genes -> "CONvsFOCUR_GM_Filtered.csv"
   9.6. Filtered "CONvsFOCURDSS_ALL.csv" -> 2way filtered (stay) -> DSS effect filtered (removed) -> **615** genes -> "CONvsFOCUR_ALL_Filtered.csv"
   9.7. Filtered "CONvsFOCURAggregate.csv" -> 2way filtered (stay) -> DSS effect filtered (removed) -> **848** genes -> "CONvsFOCURaggFiltered.csv"
   9.8. Filtered "CONvsFOCURmutual.csv" -> 2way filtered (stay) -> DSS effect filtered (removed) -> **465** genes -> "CONvsFOCURmutualFiltered.csv"

10. **Analysis combinations:**
   10.1. **(A) no DSS treatment CO control diet (CO-con) vs FO + curcumin combination** (differentially expressed on FO_DSS, CO_DSS and CO_CUR_DSS, but not differentially expressed on FO_CUR_DSS)
      10.1.1. G&M – **5**– "A_GM.csv"
      10.1.2. G&M filtered – **0**
      10.1.3. All flagged - **5**- "A_ALL.csv"
      10.1.4. All flagged filtered - **0**
      10.1.5. Based on 6.3, 7.3, 8.3, 9.3. – **6** - "A_GM_ALL_Aggregate.csv"
      10.1.6. Based on 6.7, 7.7, 8.7, 9.7. – **0**
      10.1.7. Based on 6.4, 7.4, 8.4, 9.4. – **4** - "A_GM_ALL_Mutual.csv"
      10.1.8. Based on 6.8, 7.8, 8.8, 9.8. – **0**
      10.1.9. All A mutual – **2/4** – A_MUTUAL.csv
      10.1.10. All A aggregated - **8** – A_AGGREGATE.csv

   10.2. **(B) CO-con vs curcumin supplementation** (differentially expressed on FO_DSS and CO_DSS, but not differentially expressed on CO_CUR_DSS and FO_CUR_DSS)
      10.2.1. G&M – **4** – "B_GM.csv"

10.2.2. G&M filtered – **2** - "B_GM_Filtered.csv"
10.2.3. All flagged - **7** - "B_ALL.csv"
10.2.4. All flagged filtered - **2** - "B_ALL_Filtered.csv"
10.2.5. Based on 6.3, 7.3, 8.3, 9.3. – **10** - "B_GM_ALL_Aggregate.csv"
10.2.6. Based on 6.7, 7.7, 8.7, 9.7. – **3** - "B_GM_ALL_Aggregate_Filtered.csv"
10.2.7. Based on 6.4, 7.4, 8.4, 9.4. – **2** - "B_GM_ALL_Mutual.csv"
10.2.8. Based on 6.8, 7.8, 8.8, 9.8. – **2** - "B_GM_ALL_Mutual_Filtered.csv"
10.2.9. All B mutual – **1/2** – B_MUTUAL.csv
10.2.10. All B aggregated - **11** – B_AGGREGATE.csv

10.3. **(C) CO-con vs FO supplementation** (differentially expressed on CO_DSS and CO_CUR_DSS, but not differentially expressed on FO_DSS and FO_CUR_DSS)
10.3.1. G&M – **7** – "C_GM.csv"
10.3.2. G&M filtered – **0**
10.3.3. All flagged - **13** - "C_ALL.csv"
10.3.4. All flagged filtered - **0**
10.3.5. Based on 6.3, 7.3, 8.3, 9.3. – **14** - "C_GM_ALL_Aggregate.csv"
10.3.6. Based on 6.7, 7.7, 8.7, 9.7. – **0**
10.3.7. Based on 6.4, 7.4, 8.4, 9.4. – **2** - "C_GM_ALL_Mutual.csv"
10.3.8. Based on 6.8, 7.8, 8.8, 9.8. – **0**
10.3.9. All C mutual – **0/2** – C_MUTUAL.csv
10.3.10. All C aggregated - **20** – C_AGGREGATE.csv

10.4. **(D) CO-con vs FO or curcumin supplementation** (differentially expressed on CO_DSS, but not differentially expressed on FO_DSS, CO_CUR_DSS and FO_CUR_DSS)
10.4.1. G&M – **16** – "D_GM.csv"
10.4.2. G&M filtered – **6** - "D_GM_Filtered.csv"
10.4.3. All flagged - **19** - "D_ALL.csv"
10.4.4. All flagged filtered - **3** - "D_ALL_Filtered.csv"
10.4.5. Based on 6.3, 7.3, 8.3, 9.3. – **20** - "D_GM_ALL_Aggregate.csv"
10.4.6. Based on 6.7, 7.7, 8.7, 9.7. – **4** - "D_GM_ALL_Aggregate_Filtered.csv"
10.4.7. Based on 6.4, 7.4, 8.4, 9.4. – **13** - "D_GM_ALL_Mutual.csv"
10.4.8. Based on 6.8, 7.8, 8.8, 9.8. – **4** - "D_GM_ALL_Mutual_Filtered.csv"
10.4.9. All D mutual – **1/3** – D_MUTUAL.csv
10.4.10. All D aggregated - **30** – D_AGGREGATE.csv

10.5. **(E)** Differentially expressed on FO_DSS, CO_CUR_DSS and FO_CUR_DSS, but not differentially expressed on CO_DSS

10.5.1. G&M – **161** – "E_GM.csv"

10.5.2. G&M filtered – **4** - "E_GM_Filtered.csv"

10.5.3. All flagged - **95** - "E_ALL.csv"

10.5.4. All flagged filtered - **9** - "E_ALL_Filtered.csv"

10.5.5. Based on 6.3, 7.3, 8.3, 9.3. – **202** - "E_GM_ALL_Aggregate.csv"

10.5.6. Based on 6.7, 7.7, 8.7, 9.7. – **11** - "E_GM_ALL_Aggregate_Filtered.csv"

10.5.7. Based on 6.4, 7.4, 8.4, 9.4. – **72** - "E_GM_ALL_Mutual.csv"

10.5.8. Based on 6.8, 7.8, 8.8, 9.8. – **3** - "E_GM_ALL_Mutual_Filtered.csv"

10.5.9. All E mutual – **3/3** – E_MUTUAL.csv

10.5.10. All E aggregated - **208** – E_AGGREGATE.csv

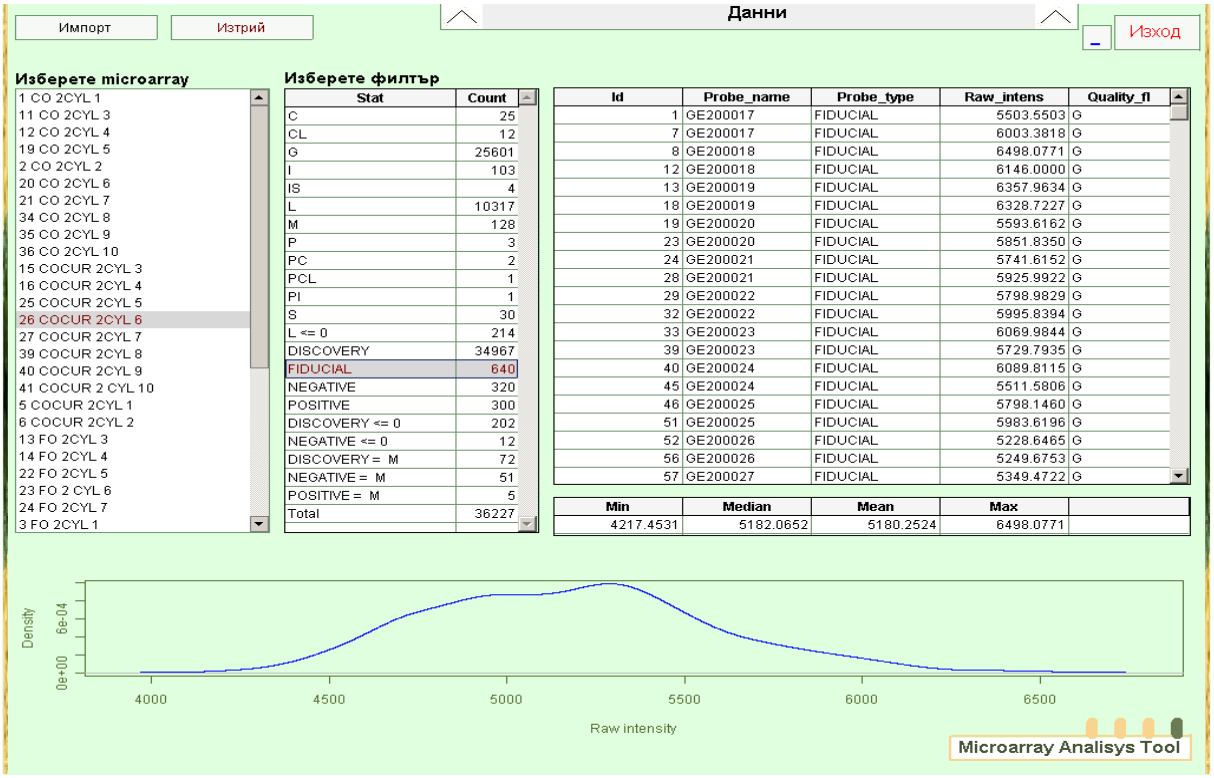
Приложение 3 – Инструкции за инсталиране

1. Изтеглете и инсталирайте последната версия на R от http://www.r-project.org/. За Windows Vista и по-нови версии на Windows инсталирайте R в директория, за която потребителят има пълни права, примерно поддиректория на главната директория.
2. От http://www.r-project.org/ изтеглете и инсталирайте следните пакети за R:
   - rscproxy
   - fdrtool
3. Изтеглете и инсталирайте R-(D)COM Interface от http://cran.r-project.org/other-software.html
4. Дазархивирайте самодезархивиращия се архив zRMicroArrayDistribute.exe в директория по ваш избор.
5. За конфигуриране на езика следвайте инструкциите в LanguageConfig.ini
6. Стартирайте основния модул на системата: zRMicroArray.exe

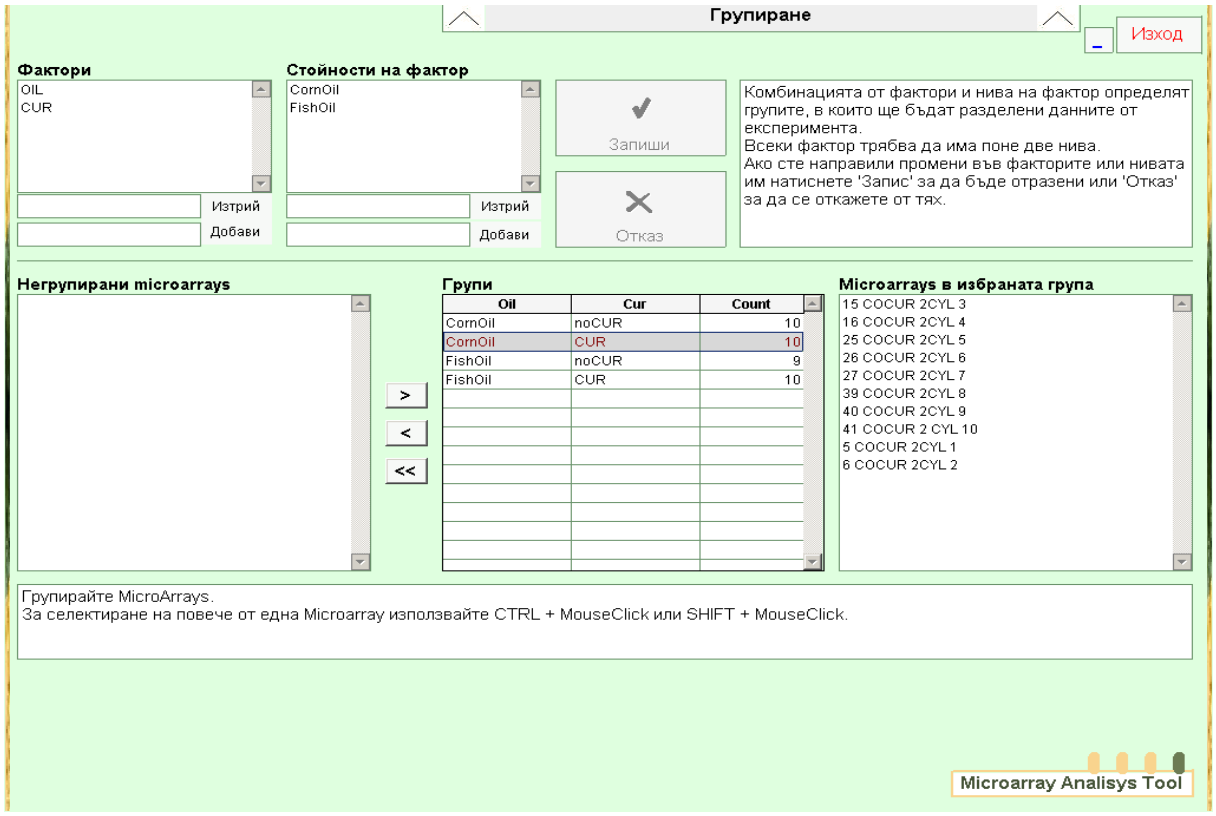
Приложение 4 – Разпечатки на екрани

- Информационен екран

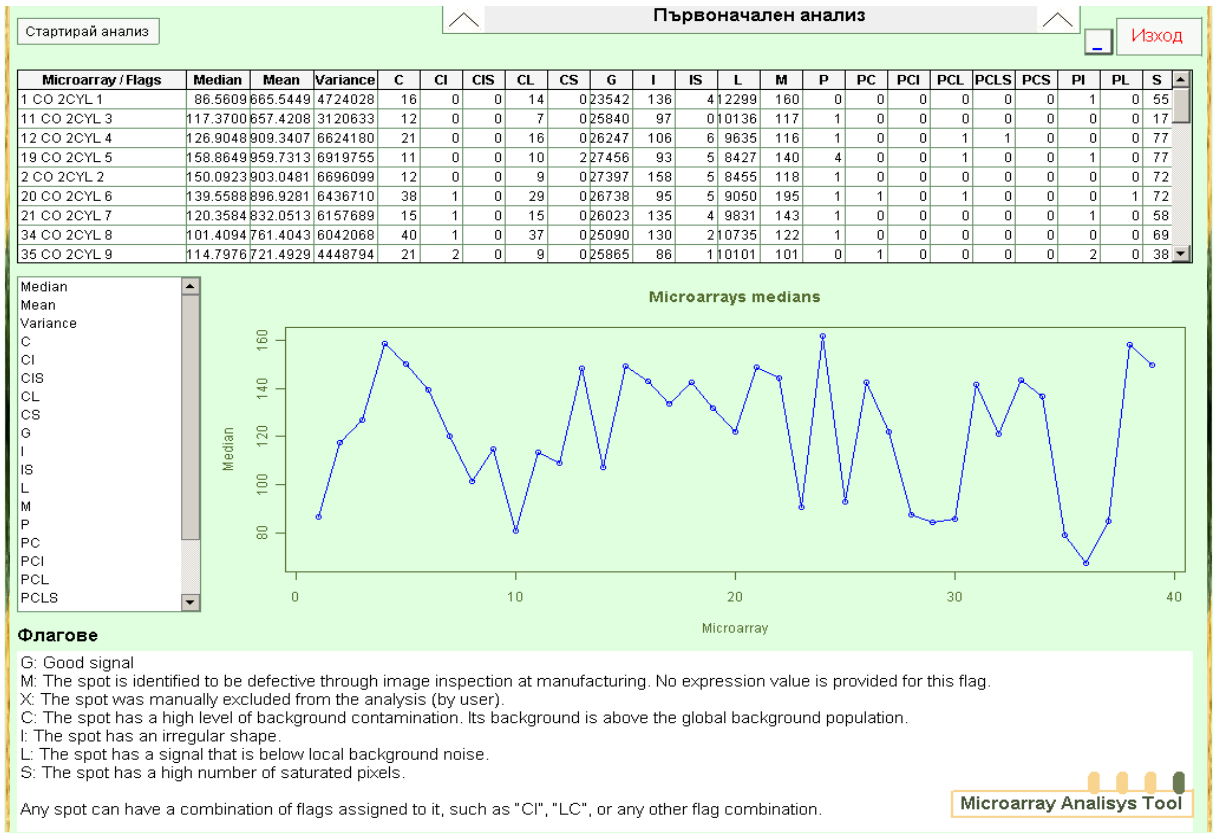
- Данни



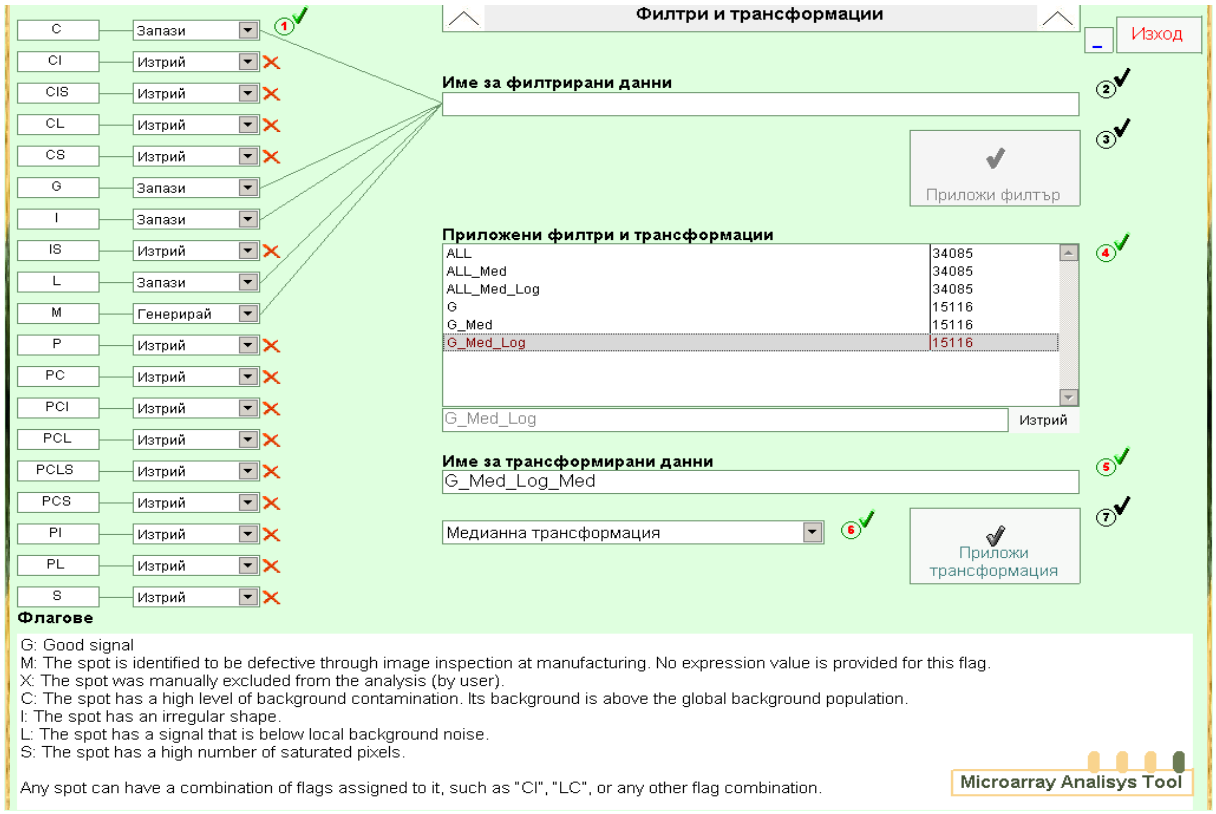
- Групиране



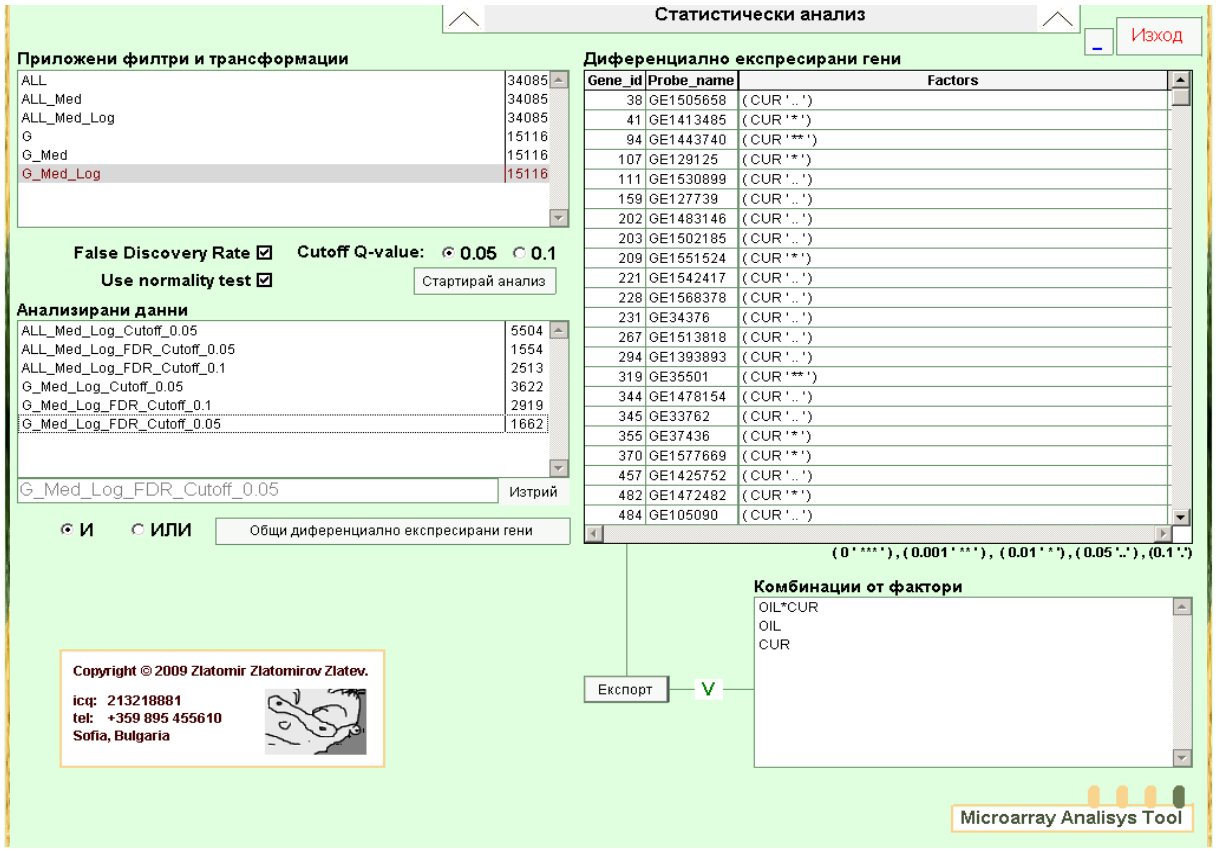
- Първоначален анализ



- Филтри и трансформации



- Статистически анализ



Приложение 5 – Терминологичен речник

Термин (български)	**Термин (английски)**	**Описание**
Генна експресия	Gene expression	Функционален генетичен продукт, синтезиран на база на ДНК
ДНК матрица (ДНК чип)	DNA microarray	Технология, позволяваща измерване на генна експресия
Поддържащи гени	Housekeeping Genes	Гени, които се експресират в относително еднакви количества, независимо от експерименталните условия и външната среда. Продуктите на тези гени изпълняват основни поддържащи функции в клетката.
Дисперсионен анализ	Analysis of variance (ANOVA)	Метод за търсене на статистически значими различия между две или повече групи нормално разпределени случайни величини.
Стойност на доверителния интервал	P-value	Вероятност със стойност между нула и едно, която отговаря на следния въпрос: Ако популацията наистина има една и съща средна стойност между групите, то каква е вероятността случайни величини да доведат до разлики между средните толкова голяма или по-голяма от наблюдаваната?
Погрешност при идентифициране	False Discovery Rate (FDR)	Резултатът от FDR са стойности на доверителния интервал (q-values) на получените от дисперсионния анализ стойности за p-values.

Приложение 6 – Експертно мнение на потенциални потребители

A comparative evaluation of zRMicroArray vs GeneSifter and GeneSpring was performed by the staff at the Genomics and Bioinformatics Facility Core, Center for Environmental and Rural Health (CERH), Texas A&M University. The table below lists the outcome of the evaluation (5 is the highest score and 1 is the lowest score). As a result of the evaluation, Dr. Robert Chapkin, the director of Genomics and Bioinformatics Facility Core, decided not to renew the core license for GeneSifter and to adopt zRMicroArray software in all of its future statistical analysis of CodeLink microarray data obtained by the core starting this September.

Questions/ Benchmarks	zRMicroArray	GeneSifter	GeneSpring
GUI	5	3	5
Ease of using	5	4	3
Speed of computing the results of a analysis	5	4	4
Graphical representation of the results	1	3	5
Cost (5 means the most expensive, 1 the least expensive)	1	4	5
Pathway analysis capable	1	4	5
Transparency of the preprocessing steps and transformations	5	3	3
Capability of processing data from different microarray platforms	3	3	4

Приложение 7 – Аудио / видео презентация на софтуера и възможностите му на български език

Файлът „PresentationBG.avi” демонстрира графичния интерфейс и възможностите на информационната система на български език.

Приложение 8 – Аудио / видео презентация на софтуера и възможностите му на английски език

Файлът „PresentationEN.avi” демонстрира графичния интерфейс и възможностите на информационната система на английски език.