

ЕРЕВАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
БИОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ
КАФЕДРА БИОФИЗИКИ

Реймерс Артур Евгеньевич

МАГИСТЕРСКАЯ РАБОТА

ТЕМА:

«Многомерный статистический подход для исследования экспрессии генов по данным микрочипов»

Научный руководитель:
проф. А. А. Чилингарян

«Допустить к защите»

Зав. кафедрой биофизики

_____ **проф. П.О.Вардеванян**

"26.05" 2004г.

Рецензент:
к.б.н. Г. В. Гюльханданян

Ереван 2004

1. Введение.

Технология нуклеотидных чипов позволяет одновременно исследовать экспрессию более 40000 генов. Большинство применяемых методов для решения задачи обнаружения генов с различной экспрессией, в двух разных клетках или тканях, основанно на одномерном подходе без учета многомерной структуры экспериментальных данных. В то время как известно, что применение ковариационных структур позволяет обнаруживать более незначительные различия.

Перед нами была поставлена задача:

- Изучить многомерный метод, предложенный в работе [3].
- Показать влияние экспериментальных ошибок на распределение экспрессии гена.
- Показать влияние экспериментальных ошибок и наличие малого количества нуклеотидных чипов на результаты отбора патологических генов.

Экспрессия гена. Экспрессией гена называются процессы транскрипции (перенос генетической информации с ДНК на РНК), сплайсинга (удаление интронов с РНК и синтеза мРНК) и трансляции (синтеза белков соответственно мРНК). Уровень экспрессии данного гена определяется измеренным количеством соответствующих транскрибированных РНК в определенном масштабе. Например, измерением может быть значение плотности флуоресцентного сигнала в виде пятна с нуклеотидного чипа.

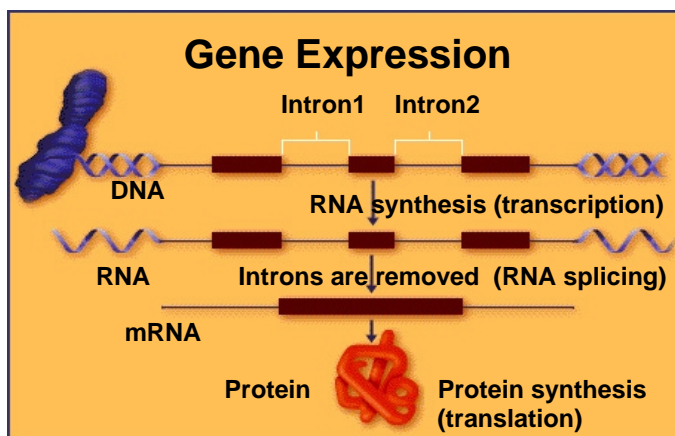


Рис. 1. Структурная схема экспрессии гена.

Нуклеотидный чип. Нуклеотидный чип представляет из себя кремниевый или стеклянный субстрат, поделенный на ячейки, и на каждую ячейку наносятся заведомо известные одноцепочечные нуклеотидные последовательности (Рис. 2). Каждая ячейка должна содержать одинаковые нуклеотидные последовательности. Количество последовательностей в каждой ячейке так же должно быть одинаково. Чем короче нуклеотидная последовательность, тем выше плотность ячеек на единице площади. Так, при длине нуклеотидной последовательности в 20 – 25 нуклеотидов количество ячеек на одном квадратном сантиметре превышает 40000. Это дает возможность одновременно исследовать экспрессию около 40000 генов.

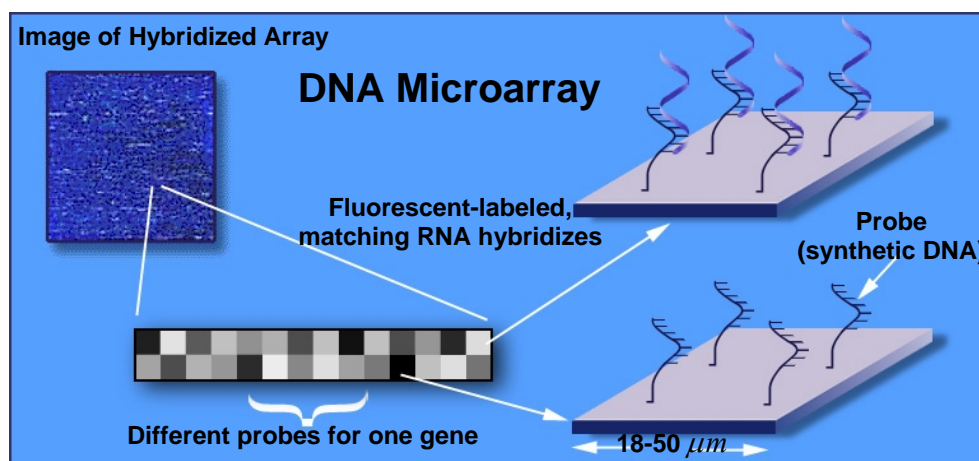


Рис. 2. Пример ячейки нуклеотидного чипа.

Данные нуклеотидных чипов можно использовать для различных исследований. Одним из таких исследований является исследование экспрессии генов в двух различных тканях (нормальной и зараженной раком) и обнаружение генов с различной экспрессией. Для этого, гены клеток зараженных раком (проверяемых) метят красным цветом, а гены нормальных (контрольных) клеток – в зеленый цвет. Потом смесь меченых генов наносят на нуклеотидный чип, в следствии чего происходит гибридизация комплементарных нуклеотидных последовательностей. Затем измеряют интенсивность флуоресцентного сигнала с каждой ячейки нуклеотидного чипа. И тем самым измеряют уровень экспрессии данного гена.

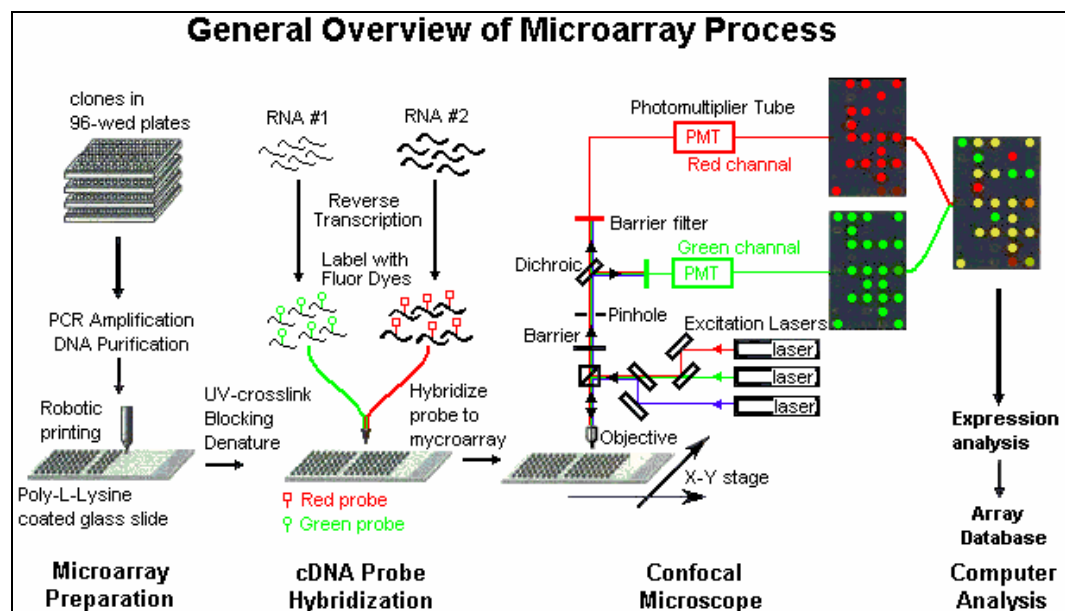


Рис. 3. Структурная схема эксперимента с использованием нуклеотидного чипа.

Методы исследования экспрессии генов. Некоторые авторы, как критерий, используют логарифм отношения красного (R) и зеленого (G) сигналов $\log_2(R/G)$ и если полученное число было больше двух или трех, то считали, что у данных двух генов различная экспрессия. Однако, известно, что уровни экспрессии различных генов имеют корреляцию, и ясно, что использование одномерного подхода не может выявить комплексы генов, ответственных за те или иные заболевания. С другой стороны, для нахождения наилучшей комбинации генов простым перебором понадобятся чрезмерные вычислительные ресурсы.

Для решения этой проблемы было предложено несколько многомерных методов ([1]–[4]). Например, в работе [1] предлагается комбинация генетического и KNN алгоритмов. В работе [2] предлагается комбинация двух алгоритмов: v -кратной перекрестной проверки и случайно выбранного метода для выделения признаков. А в работе [3] предлагается метод с использованием расстояния Махаланобиса в качестве функции качества для нахождения наилучшей комбинации генов, для которых экспрессия генов отличается наибольшим образом. Ну, а в работе [4] предлагается новый RFR (recursive feature replacement) метод, в котором используется техника SVM (support vector machines).

2. Вариабельность нуклеотидных чипов.

Чтобы показать, как влияет вариабельность между нуклеотидными чипами на распределение экспрессии одного гена, мы воспользовались программой *CorrSimulation* [5], моделирующей экспериментальные данные нуклеотидных чипов.

Чем обусловлена эта вариабельность?

Во-первых, при идеальных экспериментальных условиях, повторяя эксперимент несколько раз, для экспрессии данного гена получим некоторое распределение (Рис. 4. (Fig.1.)). Это связано с биологической вариабельностью. Кроме этого вариабельность между нуклеотидными чипами обусловлена процессом маркировки пробы и изменчивостью экспериментальных условий. Эффекты, из-за которых происходят эти ошибки, обычно делят на 3 так называемые группы [6]: «*Slide effect*», «*Dye effect*» и «*Background effect*»

«*Slide effect*». Вариабельность между нуклеотидными чипами возникает уже на стадии их изготовления, потому что невозможно на каждую ячейку чипа нанести одинаковое количество нуклеотидных последовательностей. Однако вариабельность возникает еще и потому, что от эксперимента к эксперименту количество меченых нуклеотидных последовательностей, гибридизирующихся к синтетическим последовательностям нуклеотидного чипа, меняется вследствие непостоянства экспериментальных условий (влажность, температура, степень гибридизации и т. д.) (Рис. 4 (Fig.3)). Эти все воздействия, влияющие на вариабельность между нуклеотидными чипами, принято называть «*Slide effect*».

«*Dye effect*». Другой фактор, влияющий на вариабельность между чипами, это процесс окраски (метки) нуклеотидных последовательностей. Дело в том, что разные флуоресцентные краски имеют различную устойчивость, а так же имеют разную эффективность «прилипания» к нуклеотидной последовательности. К тому же со стороны сканера они регистрируются с разной эффективностью (Рис. 4 (Fig.4)). Факторы, связанные с окраской нуклеотидных последовательностей, влияющие на вариабельность между чипами, принято называть «*Dye effect*».

И третий эффект, «*Background effect*», связан с тем, что у самого сканера есть свой фон, который накладывается при сканировании нуклеотидного чипа.

С точки зрения статистики ошибки можно разделить на две категории: это случайные и систематические ошибки. Легко понять, что «*Background effect*» вносит систематические ошибки, которые здесь мы не будем рассматривать, поскольку их можно относительно легко учесть, сканируя пустой нуклеотидный чип, получив таким образом только фон. А случайные ошибки («*Slide effect*» и «*Dye effect*») приводят к ошибкам в диагностике.

На Рис. 4 (Fig.1) показан тот случай, когда экспрессии соответствующего гена нормальной и раковой клетки не отличаются. И, естественно, мы должны были получить одинаковые распределения. На рис. 4 (Fig.2) показан случай, когда экспрессия гена контрольной и проверяемой клеток различна. В нашем случае у раковой клетки экспрессия в два раза меньше (это видно по средним значениям). И уже теперь видно что, пользуясь критерием $\log_2(R/G) > 2$, мы не выявим случай с различной экспрессией. А Рис. 4 (Fig.5) показывает, что при учете «*Slide effect*» и «*Dye effect*», заметить различие экспрессии гена в нормальной и раковой клетках практически не возможно, в то время как в действительности у раковой клетки экспрессия гена в два раза ниже. На каждом рисунке так же показаны среднеквадратические ошибки и расстояние Махаланобиса (R_m) между двумя выборками. И мы видим: чем больше распределения накладываются друг на друга, тем меньше расстояние Махаланобиса.

Gene Expression Distribution(1000 samles)

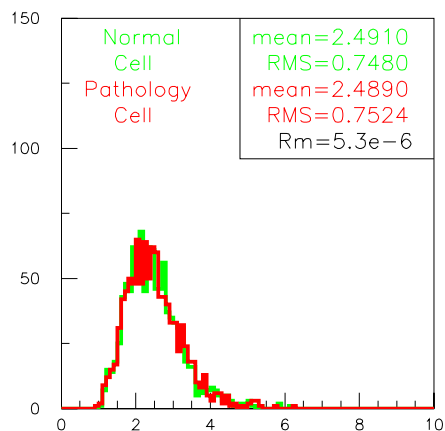


Fig.1 Without Variation without pathology

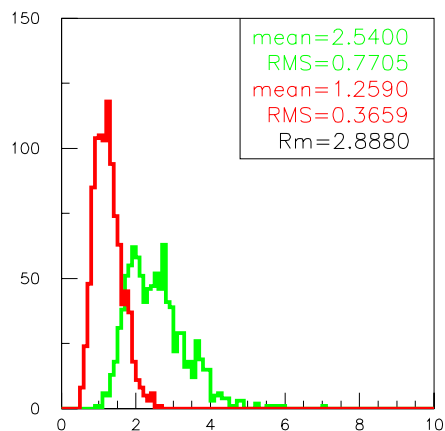


Fig.2 Without Variation with pathology

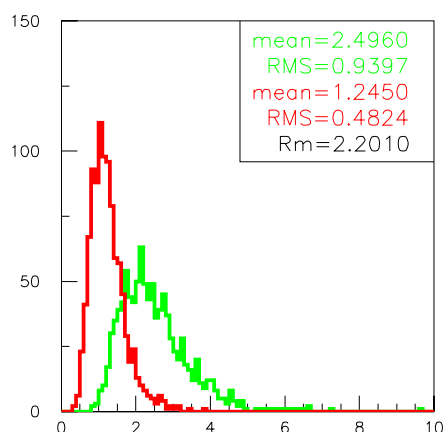


Fig.3 With Slide Effect

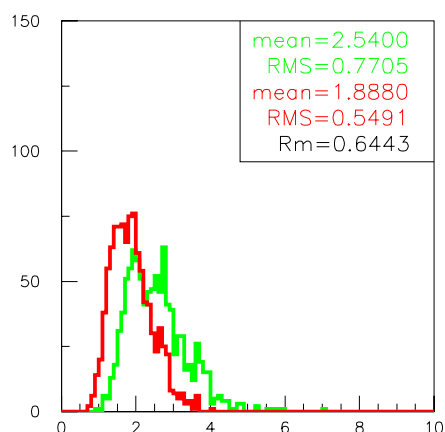


Fig.4 With Dye Effect

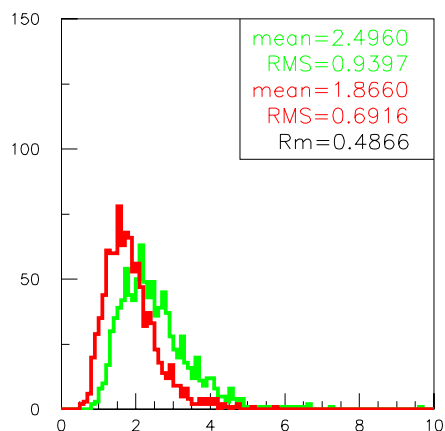


Fig.5 With All Effect

Рис. 4. Воздействие различных эффектов на распределение экспрессии гена.

3. Улучшенный метод случайного поиска для исследования экспрессии генов.

Обычно, алгоритм случайного поиска применяют для нахождения оптимальных решений в многомерном пространстве [7]. В работе [3] предлагается измененный вариант этого алгоритма. Суть его заключается в том, чтобы применить алгоритм случайного поиска некоторое количество раз, а затем объединить полученные результаты для нахождения оптимального набора генов, у которых экспрессия генов различна. Для предотвращения так называемого «overfitting» продолжительность локального случайного поиска должна быть не очень длительна. В данном алгоритме могут применяться как параметрические так и не параметрические функции качества, о которых будет сказано в следующей главе. Предложенный в работе [3] метод состоит из следующих шагов:

1. Случайным образом выбирается подмножество N_{subset} генов из общего количества генов N_{all} .
2. Оценивается функция качества для выбранных N_{subset} генов.
3. Выбирается новое подмножество N_{subset} генов путем случайной подмены одного или более генов из уже имеющегося подмножества.
4. Оценивается функция качества для нового подмножества N_{subset} генов. Если значение функции качества уменьшилось, тогда возвращаются к предыдущему подмножеству N_{subset} , в противном случае запоминают новое подмножество N_{subset} .
5. Шаги 3 и 4 повторяются N_{iter} раз, а затем запоминается соответствующее наилучшее подмножество N_{subset} генов.
6. Повторяются шаги 1 – 5 N_{cycle} раз.
7. Проводится оценка полученных N_{cycle} подмножеств генов для определения набора генов с различной экспрессией.

За нулевую гипотезу H_0 принимаем то, что сравниваемые наборы генов не отличаются в экспрессии. Следовательно, если верна нулевая гипотеза H_0 , тогда средняя частота появления различных генов должна быть одинакова и равна $N_{cycle} \cdot N_{subset} / N_{all}$ в N_{cycle} подмножествах, полученных руководствуясь пунктами 1–6. Пользуясь формулой Бернулли для малых чисел, либо теоремой Пуассона для больших, можно подсчитать, чему будет равна вероятность того, что частота появления гена будет примерно в 2 раза превышать среднее значение. При значениях $N_{cycle} = 1000$, $N_{subset} = 5$ и $N_{all} = 1000$, которые в дальнейшем используются в данной работе, эта вероятность будет равна ≈ 0.02 .

В противном случае, если не верна гипотеза H_0 , гены, имеющие большое различие в экспрессии, должны встречаться во многих N_{subset} подмножествах. Поэтому о вырожденности генов можем судить по частоте возникновения того или иного гена в N_{cycle} подмножествах.

Надо отметить, что размер N_{subset} подмножества генов ограничен количеством нуклеотидных чипов (то есть размерами выборок), поэтому N_{subset} значительно меньше N_{all} . N_{subset} так же зависит еще и от используемой функции качества. Стоит заметить, что в предлагаемом алгоритме случайного поиска обнаруживаемое количество генов с различной экспрессией не ограничивается размером N_{subset} подмножества генов. Таким образом, делая поиски малых локальных N_{subset} подмножеств, мы можем найти большой набор генов с различной экспрессией.

Число N_{iter} повторений (шагов 3 и 4) определяет границу «overfitting» - а. Оно не может быть очень малым, поскольку тогда гены с различной экспрессией не будут находиться. Но при очень большом значении N_{iter} алгоритм все время будет находить одни и те же гены из – за «overfitting».

Что касается N_{cycle} , то, именно, повторения шагов 1 – 5 позволяют нам увидеть вариабельность алгоритма случайного поиска. И это число ограничено только возможностями вычислительных ресурсов.

4. Функции качества для алгоритма случайного поиска.

Функция качества должна определять количественное различие между экспрессией генов двух сравниваемых клеток или тканей. В этой связи функции качества могут быть разными, но не надо забывать, что, если мы хотим учесть корреляцию между уровнями экспрессий генов, то наша функция качества должна содержать корреляционную структуру.

Известно, что вообще, параметрические методы более мощные чем не параметрические методы, поскольку параметрические методы включают в себя некоторую информацию о модели (например, закон распределения данных). Однако, это говорит еще и о том, что параметрические методы не совсем «универсальны», потому что чувствительны к малейшим отклонениям от модели. В случае данных нуклеотидных чипов, выборка которых мала (всего около 20 чипов), использование параметрической функции качества может привести к неустойчивым результатам, в то время как не параметрический метод случайного поиска может быть более устойчивым. В работе [3] в качестве параметрической функции качества предлагается использовать расстояние Махаланобиса

$$R_{Mah}^2 = (v - u)^T \left(\frac{\Sigma_u + \Sigma_v}{2} \right)^{-1} (v - u),$$

где v и u средние значения выборок, а Σ_v и Σ_u ковариационные матрицы выборок.

Так же предлагается расстояние Бхатачария

$$R_{Bha}^2 = \frac{1}{8} R_{Mah}^2 + \frac{1}{2} \ln \frac{|(\Sigma_u - \Sigma_v)/2|}{\sqrt{|\Sigma_u| |\Sigma_v|}}.$$

В сочетании с различными функциями качества можно использовать различные процедуры корректировки данных (понижение фона, нормировка, средне – логарифмическое и основанное на последовательности корректировки).

5. Моделирование и анализ экспериментальных данных.

Для начала мы решили повторить моделирование и анализ данных с использованием таких же параметров для N_{subset} , N_{iter} и N_{cycle} , как в работе [3]. Для этого мы воспользовались программой *CorrSimulation* [5], моделирующей экспериментальные данные нуклеотидных чипов.

Мы так же взяли 1000 генов, которые разделили на 50 кластеров (подмножества), содержащих одинаковое количество генов (20 генов в каждом кластере). В один из кластеров (для определенности первый) вводим патологию с помощью коэффициента d , принимающего случайные значения из логарифмически нормального распределения со средним значением 1 и стандартным отклонением 0.5. Корреляционная структура в двух сравниваемых наборах генов одинакова. Количество нуклеотидных чипов взяли 20. Параметры для случайного поиска взяли следующие $N_{cycle}=10000$ и $N_{subset}=5$, а в качестве функции качества взяли расстояние Махаланобиса.

На рисунке 5 показаны результаты для двух значений N_{iter} : 1000 (слева) и 100,000 (справа). Рис. 5 (Fig.1, Fig.2) показывает частоту возникновения номера последней наилучшей итерации, то есть количество итераций, после которых не найдено более лучшей комбинации. Эти две гистограммы показывают, что количество итераций $N_{iter}=1000$ не совсем достаточно для достижения глобального максимума, а количество $N_{iter}=100000$ более чем достаточно для этого.

Рис. 5 (Fig.3, Fig.4) показывает распределение расстояния Махаланобиса для наилучшей комбинации N_{cycle} подмножеств. Эти гистограммы иллюстрируют сказанное в предыдущем абзаце, только по-другому. Дело в том, что при $N_{iter}=1000$, распределение расстояния Махаланобиса имеет одномодальное распределение с большой вариабельностью. Это значит, что мы исследуем большое количество локальных максимумов. В то время как при $N_{iter}=100000$ в распределении расстояния Махаланобиса появляется дискретизация, а, так же, около половины N_{cycle} событий расстояния Махаланобиса достигают своего максимального значения, то есть у нас появляется проблема «overfitting».

На рисунке 5 (Fig.5, Fig.6) показана частота возникновения первых 20 генов. На этих рисунках так же видна проблема «overfitting», поскольку при $N_{iter}=1000$ количество генов, чьи частоты возникновения в N_{subset} подмножествах выше указанного критерия ($N_{cycle} \cdot N_{subset} / N_{all}$) – все 20 из 20. А при $N_{iter}=100000$ – их число уменьшается до 13.

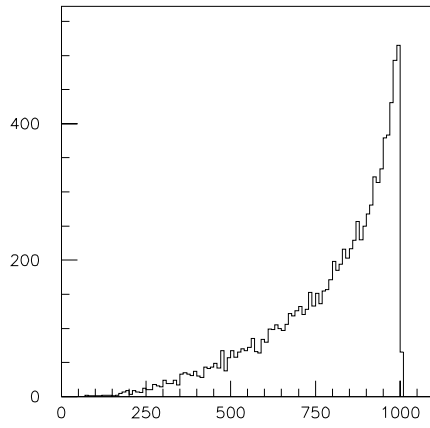


Fig.1 Last best iteration for $N_{iter}=1000$

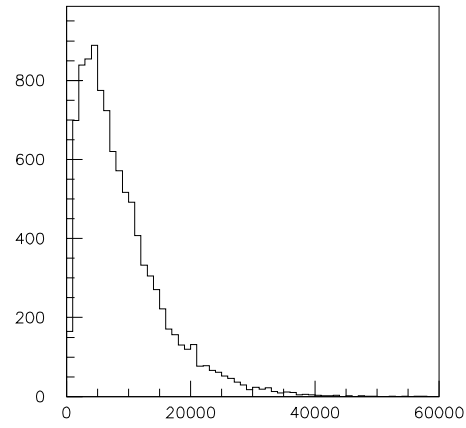


Fig.2 Last best iteration for $N_{iter}=100000$

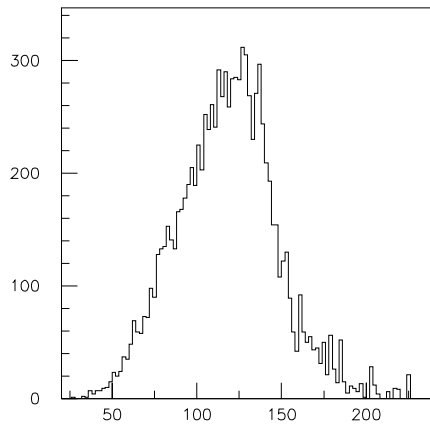


Fig.3 Mahalanobis Distance for $N_{iter}=1000$

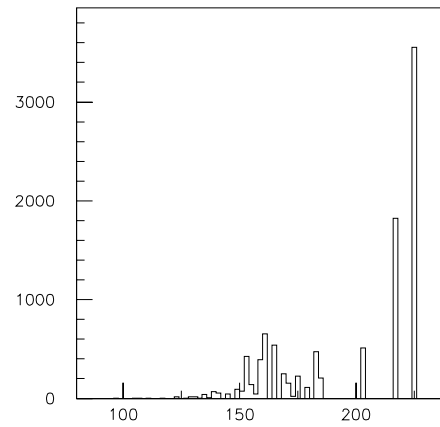


Fig.4 Mahalanobis Distance for $N_{iter}=100000$

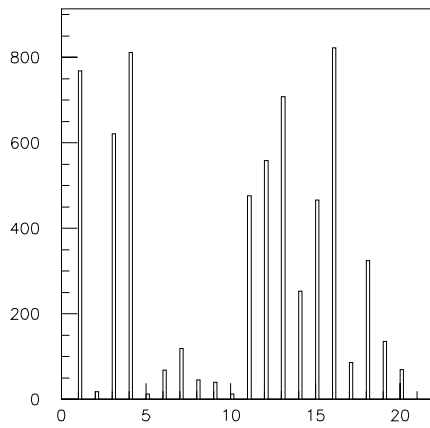


Fig.5 Frequencies of genes for $N_{iter}=1000$

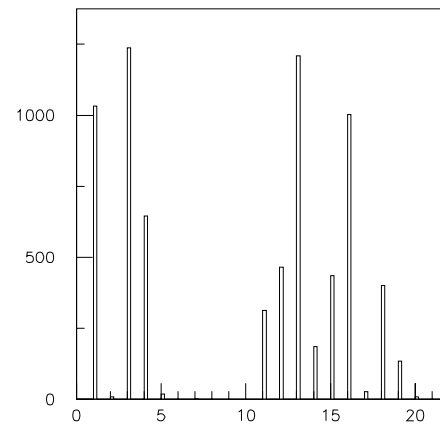


Fig.6 Frequencies of genes for $N_{iter}=100000$

Рис. 6. Сравнение отбора генов при $N_{iter}=1000$ и $N_{iter}=100000$. Fig.1 Fig.2-частота возникновения номера последней наилучшей итерации, Fig.3 Fig.4-распределение расстояния Махаланобиса для наилучшей комбинации N_{cycle} подмножеств, Fig.5 Fig.6- частота возникновения первых 20 генов.

Для того, чтобы показать, как влияет количество итераций N_{iter} на отбор патологических генов мы построили зависимости Байесового риска от значения «Cutoff» (критерия отбора) для разных значений N_{iter} (100, 200, 500, 1000, 5000, 10000).

Как известно, Байесов риск это – сумма I и II – го рода ошибок

$$Er_{Bayes} = I_{error} + II_{error} = \frac{N_{includ_NP}}{N_{all_NP}} + \frac{N_{notinclud_P}}{N_{all_P}}$$

За I род ошибки мы принимаем отношение количества не патологических генов, но зарегистрированных как патологические, к общему количеству не патологических генов (N_{includ_NP}/N_{all_NP}). В качестве II рода ошибки мы принимаем отношение количества патологических генов, но не зарегистрированных как патологические, к общему количеству патологических генов ($N_{notinclud_P}/N_{all_P}$).

Рисунок 6 (Fig.1, Fig.2) показывает зависимость Байесового риска от значения «Cutoff» для разных значений N_{iter} . Рисунок 6 (Fig.1), показывающий результаты для 20 нуклеотидных чипов, говорит о том, что не надо даже $N_{iter}=1000$ итераций. Для достижения минимума Байесового риска их количество должно быть около 200, как и значение «Cutoff». А, так же, этот рисунок иллюстрирует возникновение «overfitting», потому что при больших N_{iter} Байесов риск резко стремится к определенному значению (плато). В нашем случае, уже при значениях $N_{iter} > 5000$, Er_{Bayes} стремится к 0.5.

По рисунку 6 (Fig.2), можно сделать те же заключения, что и в предыдущем абзаце, но, делая сравнение с рисунком 6 (Fig.1), мы видим воздействие числа нуклеотидных чипов на Байесов риск. А, именно, при количестве итераций $N_{iter}=100, 200, 500$ и значениях «Cutoff» приблизительно от 200 до 500, Байесов риск равняется нулю. То есть мы находим только 20 патологических генов. Это говорит о том, что уже при 40 нуклеотидных чипах применение параметрической функции качества приводит к стабильным результатам.

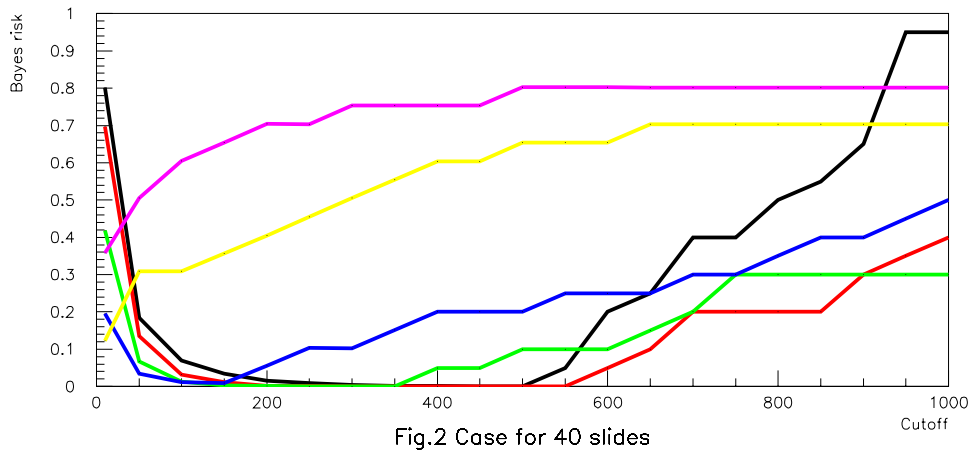
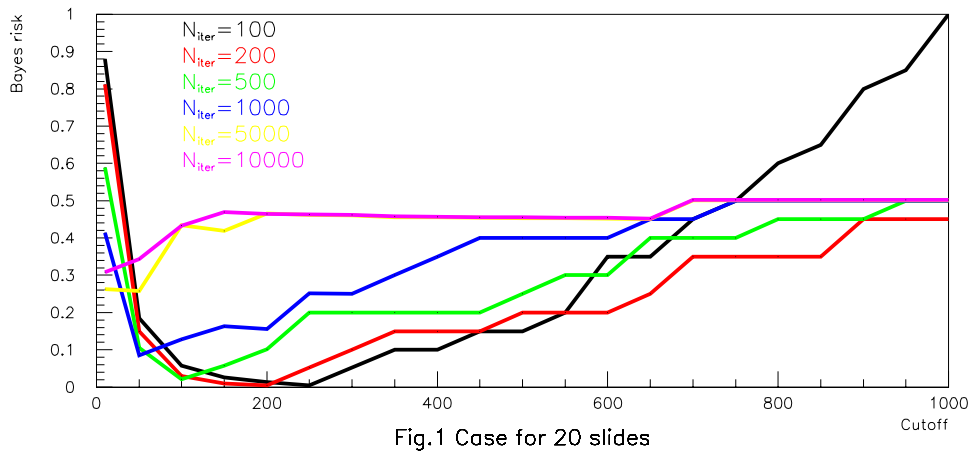


Рис. 6. Иллюстрация зависимости Байесового риска от значения «Cutoff».

Заклучение

На основе проделанной работы мы заключили, что алгоритм случайного поиска, действительно, очень эффективен, учитывая тот факт, что при количестве 40 нуклеотидных чипов и при довольно большом динамическом диапазоне значений для «Cutoff»(от 200 до 500) и довольно малом диапазоне значений для N_{iter} (от 100 до 500) отбор патологических генов 100 %. И этот алгоритм по праву носит название «алгоритм со случайным стартом и ранней остановкой».

В заключении выражаю благодарность научному руководителю проф. Чилингаряну А. А. за предложенную тему и помощь в выполнении работы. А так же выражаю благодарность за помощь в выполнении работы доктору А. Варданяну и соискателю Н. Геворкяну.

Ссылки.

1. Li L., Weinberg C.R., Darden T.A. and Pedersen L.G.: *Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method.*– Bioinformatics, 2001, Vol. 17, pp. 1131 – 1142.
2. Szabo A., Boucher K., Carroll W.L., Klebanov L.B., Tsodikov A.D. and Yakovlev A.Y.: *Variable selection and pattern recognition with gene expression data generated by the microarray technology.*– Math. Biosci., 2002, Vol. 176, pp. 71 – 98.
3. Chilingaryan A., Gevorgyan N., Vardanyan A., Jones D. and Szabo A.: *A multivariate approach for selecting sets of differentially expressed genes.*– Math. Biosci., 2002, Vol. 176, pp 59 – 69.
4. Fajarewicz K., Kimmel M., Rzeszowska – Wony J. and Swierniak A.: *A note on classification of gene expression data using support vector machines.*– J. Biol. Syst., 2003, Vol. 11, No. 1, pp. 43 – 56.
5. Aniko Szabo. *SIMULATING CORRELATED MICROARRAY DATA.* University of Utah, Department of Oncological Sciences and Huntsman, Cancer Institute.
6. George C. Tseng, Min-Kyu Oh, Lars Rohlin, James C. Liao and Wing Hung Wong: *Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects.*– Nucleic Acids Reserch, 2001, Vol. 29, No 12, pp. 2549 – 2557.
7. Zhigljavsky A.A.: *Theory of global random search.*– Mathematics and its Applications (Soviet Series), 1991, Vol. 65, Kluwer Academics, Dordrecht.