МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Механико-математический факультет

Кафедра программирования

Магистерская диссертация

БАРАМ Евгений Григорьевич

Экспертная система определения заболевания по хроматограмме образца сыворотки крови

Научные руководители
к.ф.-м.н, доцент
Ф. А. Мурзин,
Д. А. Рябичев

Новосибирск 2012

СОДЕРЖАНИЕ

1. ВВЕДЕНИЕ

Достижения в области информационных технологий, наблюдаемые за последние 20 лет, вызвали прогресс во многих областях науки, в частности, в аналитической химии, и особенно в тех ее разделах, которые связаны с обработкой большого объема экспериментальных данных. Одним из таких разделов является жидкостная хроматография — метод разделения веществ в растворе, который впервые ввел в практику М. С. Цвет в 1903 году [1]. Суть метода, показанная на рис. 1, заключается в следующем: в верхнюю часть *хроматографической колонки,* представляющей из себя трубку, наполненную мелкодисперсным адсорбентом, помещают небольшую порцию раствора анализируемого образца и промывают колонку подходящим растворителем. Важно чтобы молекулы компонентов образца в данном растворителе быстро адсорбировались и десорбировались с поверхности сорбента. В этом случае молекулы каждого типа будут передвигаться по колонке в виде узких концентрационных зон со скоростью, обратно пропорциональной силе адсорбции. Очевидно, что если сила взаимодействия адсорбата с адсорбентом для молекул разных веществ будет различной, то и скорости движения зон этих веществ будут различаться, т.е. вещества, проходя через колонку, разделяться.

Скорость движения зоны вещества зависит от скорости движения растворителя (подвижной фазы, элюента), от химического строения адсорбента и вещества, от состава элюента, от температуры.

Зависимость величины концентрации вещества вдоль зоны в идеальном случае описывается уравнением Гаусса:

$$C(V) = h \cdot e^{-0.5\left(\frac{V-V_R}{\sigma}\right)^2} \qquad (1)$$

Наблюдаемый *хроматографический пик*, соответствующий *хроматографической зоне,* изображен на рис. 2. Если измерять концентрацию веществ в растворе на выходе колонки, то мы получим кривую, которая называется *хроматограммой*.

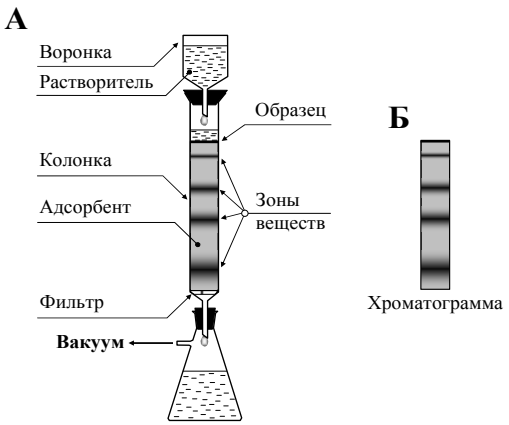


Рис. 1. Хроматограф М.С.Цвета [6].

А — схема хроматографа
Б — хроматограмма М.С.Цвета

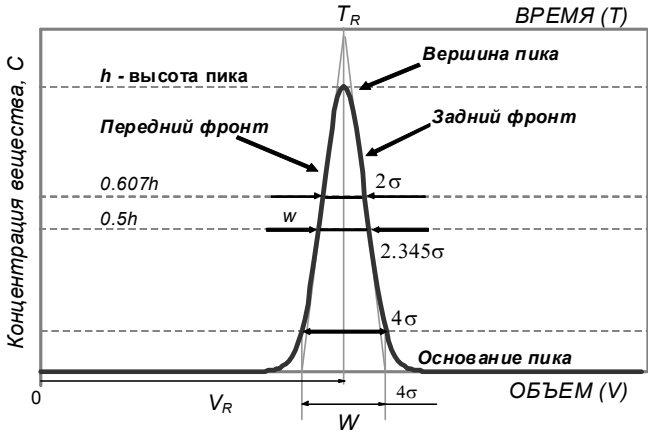


Рис. 2. Хроматографический пик.

σ — стандартное отклонение
***h* —** высота гауссова пика

Таким образом, хроматограмма является функцией зависимости концентрации вещества в растворе от объема, пропущенного через колонку растворителя или от времени. Каждому веществу на хроматограмме соответствует свой пик. Начало координат соответствует моменту ввода пробы (образца) в колонку. Абсцисса вершины пика называется *объемом удерживания вещества* (V_R), величина которого определяется химическим строением этого вещества, составом *подвижной фазы,* свойствами адсорбента (*неподвижной фазы*) и температурой. Когда говорят о *времени удерживания* (T_R), то имеют в виду, что , где F — скорость потока растворителя через колонку. Мерой количества вещества, введенного в колонку, является площадь хроматографического пика, равная

$$S = \int_{V_1}^{V_2} C(V)dV \qquad (2)$$

где V_1 и V_2 – «начало» и «конец» хроматографического пика.

Концентрация вещества в подвижной фазе, вытекающей из колонки (*элюат*), измеряется с помощью *детектора*, представляющего собой специальное устройство с проточной измерительной ячейкой, выходной сигнал которого пропорционален концентрации вещества в растворе. Устройство детектора может быть основано на многих физико-химических

принципах, но мы рассмотрим только *фотометрический детектор*, который применяется в хроматографе «Милихром А-02» (ЗАО Институт хроматографии «ЭкоНова», г. Новосибирск), который использовался в данной работе для разделения смесей веществ.

Принцип работы фотометрического детектора показан на рис. 3:

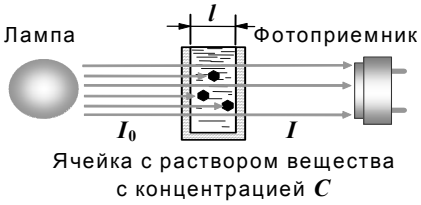


Рис. 3. Схема фотометра.

Пучок света от источника интенсивностью I_o проходит через ячейку с раствором вещества с концентрацией C и попадает на фотоприемник. Так как часть света поглощается веществом, интенсивность света на выходе $I < I_0$. Этот процесс описывается уравнением Бугера—Ламберта—Бера:

$$I = I_0 \cdot 10^{-\varepsilon_\lambda \cdot C \cdot l} \qquad (3)$$

где ε_λ — коэффициент (коэффициент *экстинкции*), характеризующий поглощение света с длиной волны λ раствором вещества с концентрацией $C = 1$ в кювете с длиной оптического пути $l = 1$. В логарифмической форме это уравнение выглядит как:

$$A_\lambda = \log\left(\frac{I}{I_0}\right) = \varepsilon_\lambda \cdot C \cdot l \qquad (3')$$

Важно, что в уравнении (3') поглощение (*оптическая плотность*) раствора вещества при длине волны λ прямо пропорционально концентрации вещества. Величину A_λ принято измерять в единицах оптической плотности (е.о.п.).

Длина волны λ фотометрического детектора выбирается, как правило такой, чтобы обеспечить необходимую чувствительность анализа и определяется из спектра поглощения раствора вещества. Спектральный диапазон детектора хроматографа «Милихром А-02» равен 190÷360 нм (УФ

область электромагнитного спектра), что позволяет регистрировать концентрацию подавляющего количества веществ. Типичный УФ-спектр показан на рис. 4:

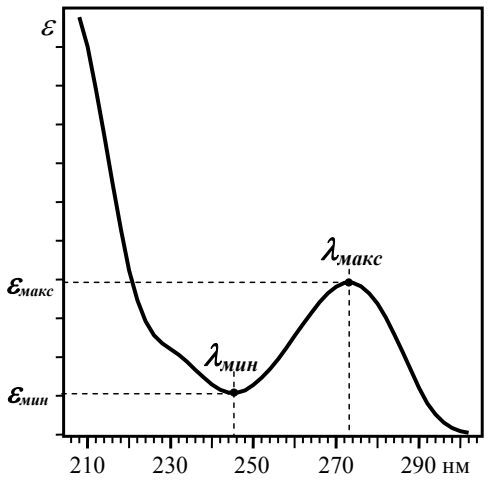


Рис. 4. УФ-спектр водного раствора кофеина.

Регистрация оптической плотности при одной длине волны является простейшим способом детектирования, при котором можно лишь определить количество вещества по площади пика, но нельзя идентифицировать это вещество по его спектральным характеристикам. Детектор хроматографа «Милихром А-02» может работать в многоволновом режиме, быстро перестраивая монохроматор по циклической программе (до восьми длин волн в цикле) так, что концентрация вещества в элюате, протекающем через фотометрическую ячейку, успевает изменится лишь незначительно. Хроматографический пик при многоволновом детектировании выглядит так, как показано на рис. 5.

Кроме того следует отметить, что существуют т.н. диодно-матричные детекторы [2]: в них несколько фотоприемников объединены в матрицу (стоящую после рассеивающей призмы) таким образом, что каждый из них фиксирует излучение на определенной длине волны. Это создает большие трудности в процессе анализа, т.к. при малой интенсивности излучения каждый приемник получает недостаточно света, что сильно повышает уровень шумов. Если же использовать более мощный источник света, то

исследуемый образец может начать разлагаться, и на хроматограмме будет показан спектр не самой пробы, а продуктов ее фотолиза. Кроме того, сильное излучение вызывает сильный нагрев матрицы и, как следствие, ее расширение, в результате чего фотоприемники смещаются и начинают фиксировать часть света на соседних длинах волн. Чтобы избежать этого, матрицы приходится жестко термостатировать (вплоть до -270°С).

Информационная ценность многоволнового детектирования заключается в том, что кроме объема удерживания для идентификации вещества можно использовать нормированные величины $R=A_{\lambda n}/A_{\lambda m}$ (*спектральные отношения*), измеренные в один момент времени, когда концентрация вещества в растворе неизменная.

Исходя из спектра вещества, легко показать, что

$$R = \frac{A_{\lambda n}}{A_{\lambda m}} = \frac{\varepsilon_{\lambda n} \cdot C \cdot l}{\varepsilon_{\lambda m} \cdot C \cdot l} = \frac{\varepsilon_{\lambda n}}{\varepsilon_{\lambda m}} = const \qquad (3)$$

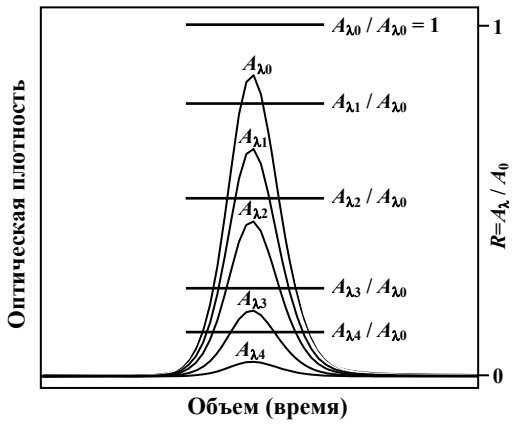


Рис. 5. Хроматографический пик, зарегистрированный при 5 длинах волн.

Таким образом, набор спектральных отношений отражает «спектральный портрет» вещества и может быть использован для идентификации вещества, а постоянство значений R вдоль хроматографического пика является критерием его «чистоты».

Если мы начали с того, что определили хроматографию как метод разделения веществ, то в современной терминологии этот метод включает в себя и измерение концентрации веществ в образце, и их идентификацию по

различным параметрам. Другими словами, хроматография отвечает на вопросы:

1. Сколько различных веществ в образце?
2. Есть ли в образце вещества $x_1, ..., x_n$ из известного нам списка?
3. Какова концентрация веществ $x_1, ..., x_n$ в исходном образце?

Для того, чтобы получить ответы на вопросы 2 и 3, образцы стандартных (*эталонных*) растворов веществ $x_1, ..., x_n$, хроматографируют в стандартных условиях, регистрируя их объемы удерживания, площади пиков и спектральные отношения, причем для каждого измеряемого параметра определяется погрешность измерения [3]. Вся эта процедура называется калибровкой хроматографа. Если калибровочные параметры стабильны во времени, то их сохраняют в виде баз данных. Анализ исследуемой пробы проводят строго в тех же условиях. Полученные данные обрабатывают, учитывая дрейф сигнала детектора, уровень его шумов, неполное разделение пиков и т.п., а затем сравнивают экспериментальные данные со стандартными, взятыми из базы данных.

2. ПОСТАНОВКА ЗАДАЧИ

Интересным и важным направлением ВЭЖХ-УФ в последние годы можно считать «стандартизацию» обзорного анализа сыворотки крови человека. После обработки хроматографической и спектральной информации становится возможным отслеживать значимые изменения в составе крови и, тем самым, проводить диагностику ряда заболеваний, включая онкологические.

Нашей задачей является разработка компьютерной системы, способной обрабатывать результаты такого анализа. В этом процессе выделяются два основных этапа: обработка сигнала детектора — сглаживание, приведение к единой временной сетке, удаление выбросов и кластеризация, и последующий анализ полученных компонентов с целью сравнения их с эталоном или идентификации по базе данных «ВЭЖХ-УФ». Идентификация компонентов может производиться по времени удерживания и спектральным отношениям, что дает возможность различить более чем 10^{11} веществ [3].

Основным назначением системы будет являться диагностика заболеваний по хроматограмме образца сыворотки крови. Результат работы будет представляться в виде степени принадлежности рассматриваемого образца к кластерам, построенным по обучающим выборкам хроматограмм сыворотки крови здоровых людей и людей, имеющих определенные заболевания.

3. ОСНОВНЫЕ ПОДХОДЫ К ПРОБЛЕМЕ

Для решения задачи сравнения двух хроматограмм в мире применяются два основных метода: изучение корреляции двух массивов данных [4] и кластеризация этих массивов [5, 6] и последующее сравнение отдельных пиков [6]. Рассмотрим эти подходы подробнее.

В первом случае на основе записанных данных вычисляется коэффициент корреляции двух хроматограмм — чаще всего применяются методы Пирсона и Стьюдента [4]. При значении коэффициента корреляции >0,99 хроматограммы считаются подобными. Коэффициент от 0,99 до 0,90 указывает на некоторое сходство, но результат следует интерпретировать с осторожностью. Значения ниже 0,9 подразумевают, что образцы различны. Данный метод достаточно прост в реализации и имеет достаточно высокую эффективность, однако позволяет ответить лишь на вопрос «идентичны представленные образцы или нет». Кроме того, если возникнет необходимость сохранения хроматограммы представленного образца в базу данных, потребуется запись всех имеющихся точек, т.е. около чисел с плавающей запятой.

Второй метод подразумевает разбиение каждой хроматограммы на отдельные пики для получения т.н. *«хроматографических отпечатков»* (*«chromatographic fingerprints»* [3]). Этот процесс намного более трудоемкий, однако имеющий два основных преимущества. Во-первых, при обработке учитывается намного больше факторов: дрейф пиков и базовой линии, пересечение двух и более пиков, и т.д. Во-вторых, метод позволяет описать *каждый* пик набором из десятка чисел, что, с одной стороны значительно сокращает объем данных, а с другой стороны дает возможность ответить на вопросы «идентичны ли представленные образцы, и если нет — то чем они отличаются», «содержится ли в представленном образце данное вещество», а при наличии базы чистых веществ и на вопрос «из каких веществ состоит представленный образец». Именно этот метод мы будем использовать в дальнейшем.

4. ФИЛЬТРАЦИЯ ШУМОВ

Первым этапом обработки полученных от хроматографа данных является фильтрация (или сглаживание) шумов. Основным препятствием в этом процессе является тот факт, что на хроматограммах всегда имеются два вида шумов: электронный шум, возникающий в результате работы АЦП детектора, и химический шум, являющийся результатом детектирования большого количества случайных примесей, неизбежно имеющихся в любой пробе.

Были реализованы несколько алгоритмов, позволяющих бороться с шумами обоих типов:

Вейвлет-преобразование Хаара

Для построения вейвлета размера N используется функция Хаара:

$$h_k(t) = \begin{cases} 2^{\frac{p}{2}}, если \frac{q-1}{2^p} \le t < \frac{q-0.5}{2^p} \\ -2^{\frac{p}{2}}, если \frac{q-0.5}{2^p} \le t < \frac{q}{2^p} \\ 0 \ иначе \end{cases} \qquad (6)$$

определенная на интервале $0 \le t \le 1$ (k=0, 1. ..., N-1). Параметр p определяется из соотношений $2^p < k$ и $2^{p+1} \ge k$, параметр q равен $k - 2^p + 1$. Матрица вейвлета строится из значений функции $h_k(t)$ при $t = \frac{m}{N}$ $(m = 0,1,...,N-1)$.

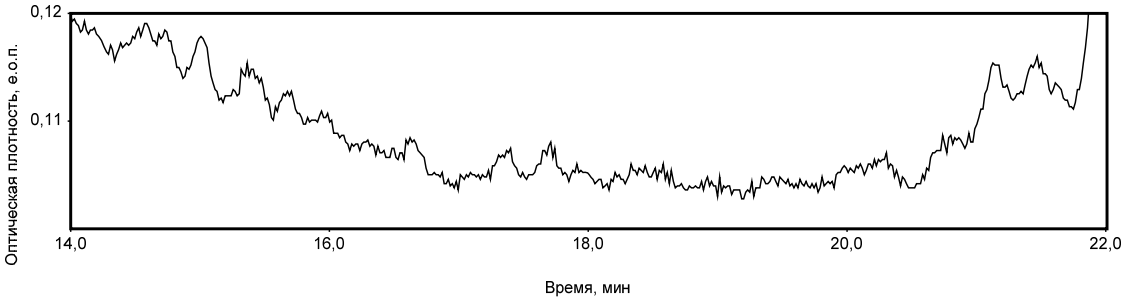
Фильтр Савицкого-Голея

Сглаживающий фильтр Савицкого-Голея позволяет снизить уровень шума не внося значительных искажений в площадь пиков [4]. Значение сглаживающей функции в точке x_m вычисляется по формуле:

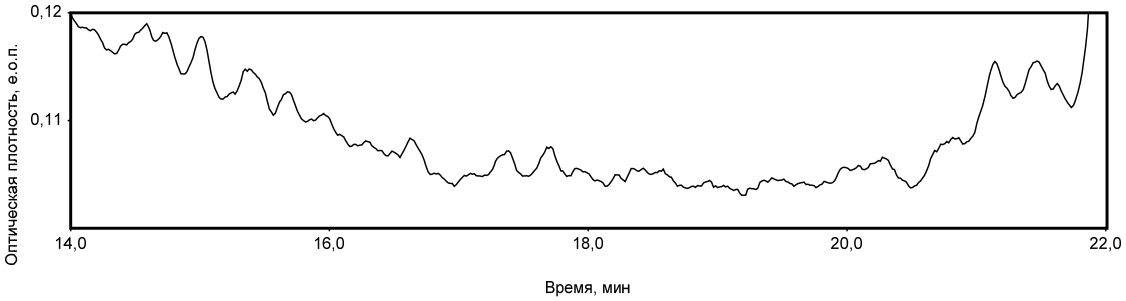
$$SG_N(x_m) = \sum_{i=-\frac{N-1}{2}}^{\frac{N-1}{2}} \frac{p_{N_i} \cdot x_{m+i}}{h_N} \qquad (7)$$

где N - ширина окна фильтра (имеет вид $2k$+1), p_N и h_N - весовые параметры: p_5={17,12,-3}, h_5=35, p_7={7,6,3,-2}, h_7=21, и т.д.

а)



б)



в)

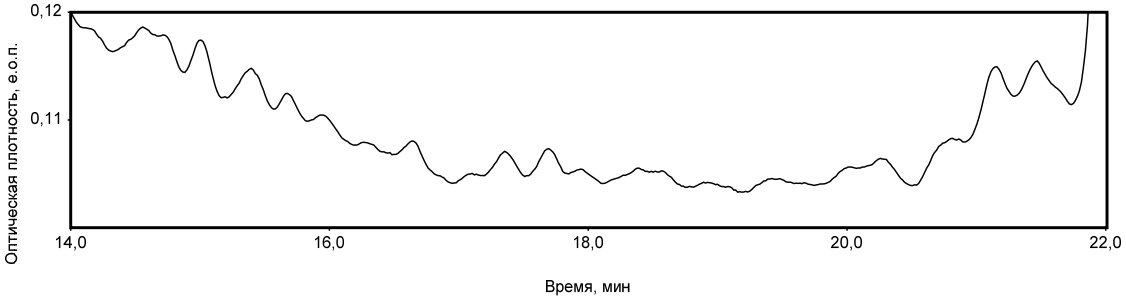


Рис. 6. *а)* нефильтрованные данные, *б)* данные, обработанные фильтром Савицкого-Голея с окном ширины 7, *в)* данные, обработанные фильтром Савицкого-Голея с окном ширины 15

Медианный фильтр

Медианный фильтр является одним из нелинейных фильтров с конечной импульсной характеристикой [7, 8]. Значения отсчетов внутри окна фильтра сортируются в порядке возрастания (убывания); и значение, находящееся в середине упорядоченного списка, поступает на выход фильтра. В случае четного числа отсчетов в окне выходное значение фильтра равно среднему значению двух отсчетов в середине упорядоченного списка. Затем окно перемещается вдоль фильтруемого сигнала и вычисления повторяются.

Также в работе используются вейвлет-преобразование Добеши-4 и сглаживающий фильтр Гаусса [9, 10]. Применяя их в различных сочетаниях удается снизить уровень шума, не внося значительных изменений в значения площадей пиков. Уровень шумов оценивается как среднеквадратичное отклонение в выборке разности сигналов (приращений) двух соседних точек, очищенной от пиков и случайных выбросов.

5. ВЫДЕЛЕНИЕ ПИКОВ

Одной из наиболее важных процедур является разбиение хроматограммы на пики. Процедура поиска параметров пиков называется *интегрированием* [8,11]. Интегрирование включает в себя определение особых точек пиков (*начало*, *конец*, *вершина*, *долина*), построение *базовой линии*, вычисление таких характеристик пиков, как *время удерживания* (т.е. время выхода вершины пика), *высота* и *площадь*. Обычно величиной, характеризующей содержание компонента в анализируемой смеси, является именно площадь пика.

Сверху пики ограничены хроматографической кривой, а снизу — базовой линией. Слившиеся пики (неразделенные у основания) объединяются в группу, где конец предыдущего пика совпадает с началом следующего (эта общая точка называется долиной). В этом случае базовая линия начинается в точке, относящейся к началу первого пика, и заканчивается в точке, относящейся к концу последнего пика в группе и считается прямой. На данном этапе для разграничения смежных пиков применяются два метода: «метод долин» (базовая линия проводится по долинам группы) и «метод перпендикуляров» (смежные пики разграничиваются вертикальной прямой, соединяющей хроматографическую кривую с базовой линией). Каждый из этих методов имеет свои достоинства и недостатки, которые следует рассмотреть более подробно.

Метод долин, с одной стороны, дает достаточно неплохое приближение, однако имеет ряд существенных недостатков. Во-первых, пик, выглядящий как плечо на склоне более высокого пика, не будет распознан. Во-вторых, может быть проигнорирована большая площадь, лежащая ниже новой базовой линии [6].

Метод перпендикуляров, в свою очередь, дает возможность выделить пики-наездники (пусть и с относительно низкой точностью) и сохранить для дальнейшего изучения всю площадь под хроматографической кривой. Однако этот метод хорошо подходит только для примерно равных по

площади пиков: так как начало второго пика оказывается в области первого, а конец первого — в области второго, то ошибка будет минимальной при наименьшей разности площадей перекрывающихся областей. Если же размер пиков сильно отличается (например, как 20:1), то большой пик будет лишь немного «испорчен» маленьким, в то время как маленький получит большой кусок площади большого. В этом случае метод долин может дать значительно более точный результат [5].

Для многоволновых же хроматограмм существует еще один метод, основанный на изучении спектральных отношений. Дело в том, что для чистых веществ оптическая плотность A при длине волны λ_1 прямо пропорциональна плотности на другой длине волны λ_2 (согласно закону Бугера—Ламберта—Бера):

$$A(\lambda_1)=k\cdot A(\lambda_2),\ k\equiv const \qquad (8)$$

Любая примесь, элюирующаяся вместе с основным пиком, вызовет отклонение линии относительного поглощения от горизонтальной прямой. Изучая поведение этой линии (см. рис. 7), можно с уверенностью сказать является ли пик чистым или представляет собой смесь нескольких веществ [3].

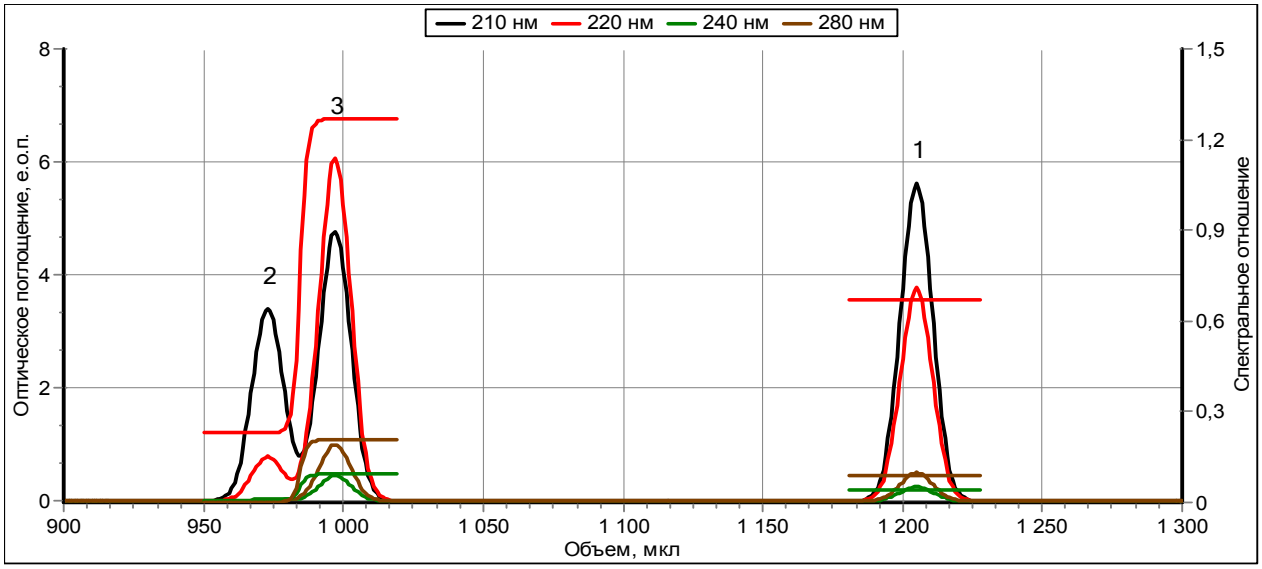


Рис. 7. Фрагмент смоделированной на компьютере хроматограммы раствора 3-х пептидов с детектированием при четырех длинах волн (210, 220, 240 и 280 нм). Справа показан *гомогенный* (чистый) пик 1 с постоянными значениями спектральных отношений вдоль всего пика. Слева — два *гетерогенных* (неразделенных) пика 2 и 3.

В последнем случае, если два пика перекрывают друг друга так, что каждый из них имеет фрагмент (начало первого и конец второго), свободные от примесей, то, зная лишь сумму их сигналов, можно решить обратную задачу и разделить пики точно:

$$A_1^1 = A_1^* - A_1^2 \tag{9}$$

$$A_1^2 = \frac{A_1^*}{k_2 - k_1} \cdot \left(\frac{A_2^*}{A_1^*} - k_1 \right) \tag{10}$$

где A_j^i — оптическая плотность i-го пика на j-ой длине волны, A_j^* — сумма сигналов на j-ой длине волны, k_i — спектральное отношение для i-го пика ($A_{\downarrow}1^{\uparrow}i / A_{\downarrow}2^{\uparrow}i$) в начале первого пика и в конце второго соответственно. Аналогично строятся выражения для трех, четырех и более перекрывающихся пиков. Этот метод носит название «факторный анализ».

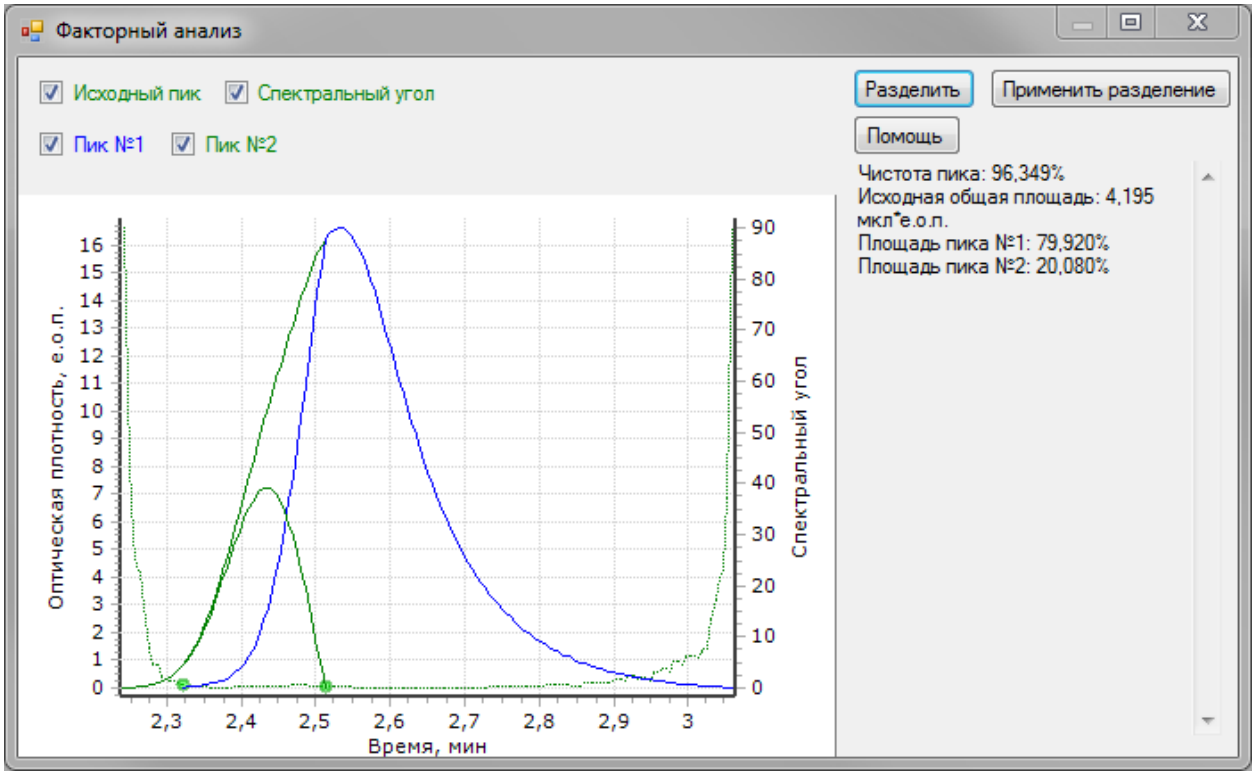


Рис. 8. Результат разделения двух перекрывающихся пиков методом факторного анализа.

Разбиение хроматограммы на отдельные пики может проводиться как в автоматическом режиме, так и вручную. При выборе автоматической разметки есть возможность задать набор параметров, описывающих

минимальный отмечаемый пик (высоту, ширину и площадь), а также начало и конец обрабатываемой области, канал для разметки и тип используемой базовой линии. Начало и конец пика определяются исходя из соотношения значения производной и уровня шума [7, 13].

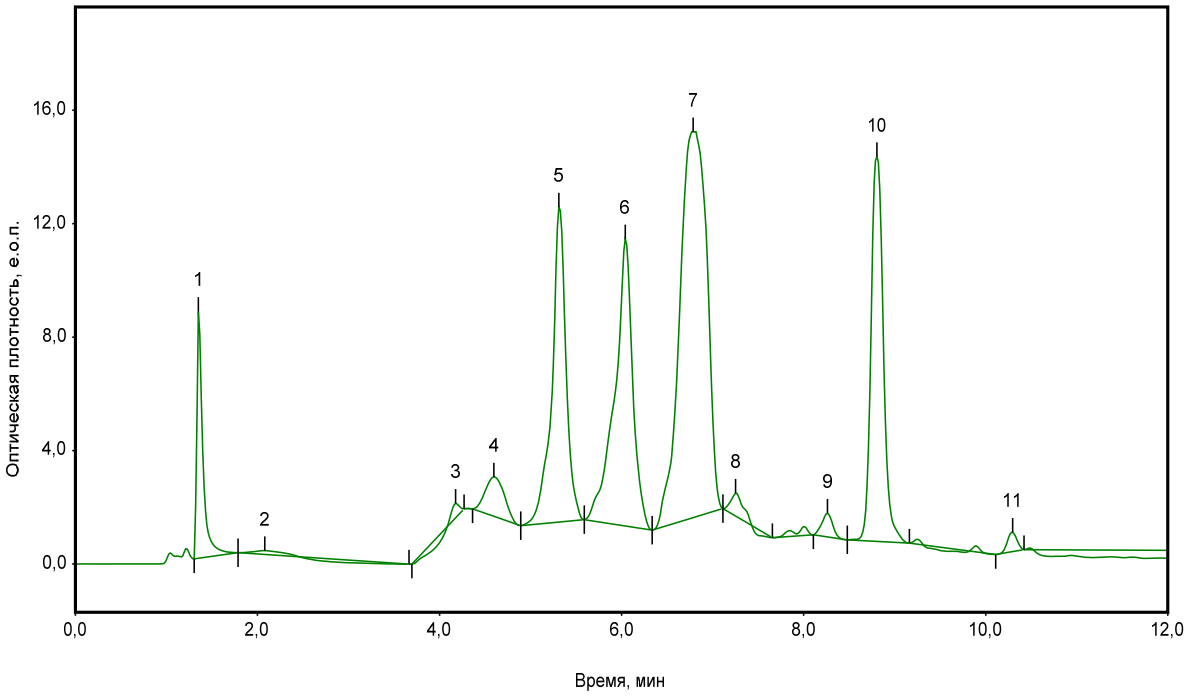


Рис. 9. Результат автоматического разбиения хроматограммы зеленого чая на пики. Для разграничения смежных пиков выбран метод долин.

В ручном режиме разбиения пользователь может самостоятельно отметить интересующие его пики, изменить тип базовой линии или внести поправки в результат автоматического разбиения.

Цель разбиения хроматограммы на пики заключается в том, чтобы перейти от работы с сигналом детектора, представляющего собой несколько сотен тысяч точек, к значительно более компактному набору объектов-пиков. Каждый такой объект является набором из следующих параметров:

- время удерживания вещества (координата вершины пика по оси абсцисс)
- площадь пика
- набор спектральных отношений

Если суть первых двух параметров очевидна, то о наборе спектральных отношений стоит сказать чуть подробнее. В введении мы уже упоминали понятие спектрального отношения — это значение $A_{\lambda n}/A_{\lambda m}$, где $A_{\lambda k}$ — значение оптической плотности на k-ой длине волны. В нашем случае мы работаем одновременно с 8 длинами волн, первую из которых считаем опорной и нормируем все оставшиеся значения по значению оптической плотности на ней. Таким образом, спектр пика представляет собой 7 значений, образующие вектор в 7-мерном пространстве.

Главной проблемой, возникающей при работе со спектром пика, является выбор момента времени, в котором будут браться искомые значения оптической плотности. В разрабатываемой системе предлагается три различных варианта:

1. Брать значения спектра *в вершине пика*. Этот метод является наиболее очевидным, однако, при высоком уровне шумов или при наличии достаточно большого количества примесей в образце, полученный спектр может быть сильно искажен.
2. Использовать *значения площадей пика* на различных длинах волн вместо значений оптической плотности. Этот метод дает неплохой результат при высоком уровне шумов (которые слабо влияют на площадь пика), но наличие примесей по-прежнему вызовет искажения в полученном спектре.
3. Брать значения оптической плотности в точке, где кривая спектральных отношений *наиболее близка к прямой*. Данный метод наиболее сложен для реализации, но дает наилучший результат даже при наличии примесей. Как упоминалось ранее (см. рис. 5), пик вещества без примесей имеет постоянные спектральные отношения. Предлагаемый метод выбирает участок пика, наиболее свободный от примесей, и выдает спектр почти чистого вещества.

Сразу скажем несколько слов о метрике для спектральных векторов, которую мы будем использовать в дальнейшем. Расстоянием между двумя спектрами мы будем называть угол, образуемый двумя их векторами в

пространстве [14, 15]. Иными словами, расстояние между спектрами a=(a_1, a_2, a_3, a_4, a_5, a_6, a_7) и b=(b_1, b_2, b_3, b_4, b_5, b_6, b_7) равно:

$$d(a,b) = \arccos\left(\frac{\sum_{i=1}^{7} a_i b_i}{\sqrt{\sum_{i=1}^{7} a_i^2} \cdot \sqrt{\sum_{i=1}^{7} b_i^2}}\right) \quad (11)$$

В данной работе полученный угол будет измеряться в градусах и обозначаться «°».

После выполнения разбиения хроматограммы появляется возможность вычислить все необходимые параметры полученных пиков. Пример отчета по такой обработке можно увидеть в Приложении на стр. 39.

6. ГРАДУИРОВКА

В системе реализована система градуировок, позволяющая вычислять концентрацию того или иного вещества в образце, основываясь на данных об объеме пробы, площади соответствующего пика и заранее заданных эталонных концентрациях.

Известно, что площадь пика линейно зависит от концентрации соответствующего вещества в образце. Аппроксимация производится с использованием метода наименьших квадратов и имеет вид

$$C = k \cdot \frac{S}{V} \quad \text{или} \quad C = k\frac{S}{V} + b$$

где C — концентрация, S — площадь пика, а V — объем пробы.

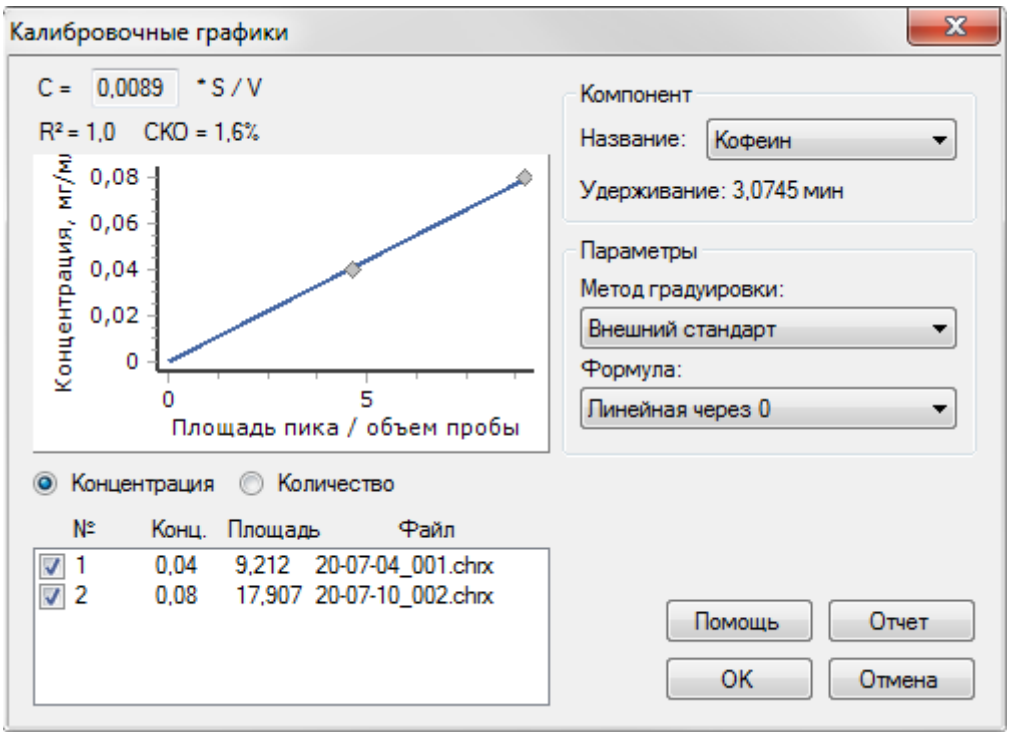


Рис. 10. Диалог выбора градуировочных параметров

В результате градуировки появляется возможность увидеть уже не относительные изменения концентраций компонентов, а абсолютные. Примеры отчетов по градуировке приведены в Приложении на стр. 36

.

7. БАЗА ДАННЫХ ПО ВЕЩЕСТВАМ

В качестве первого шага к решению проблемы идентификации образца реализован механизм поиска соответствующего компонента по базе спектральных данных. Эта база содержит в себе данные по 500 веществам, для каждого из которых были измерены спектральные отношения, время либо объем удерживания и удельные площадь пика и концентрация для автоматической градуировки (см. рис. 11).

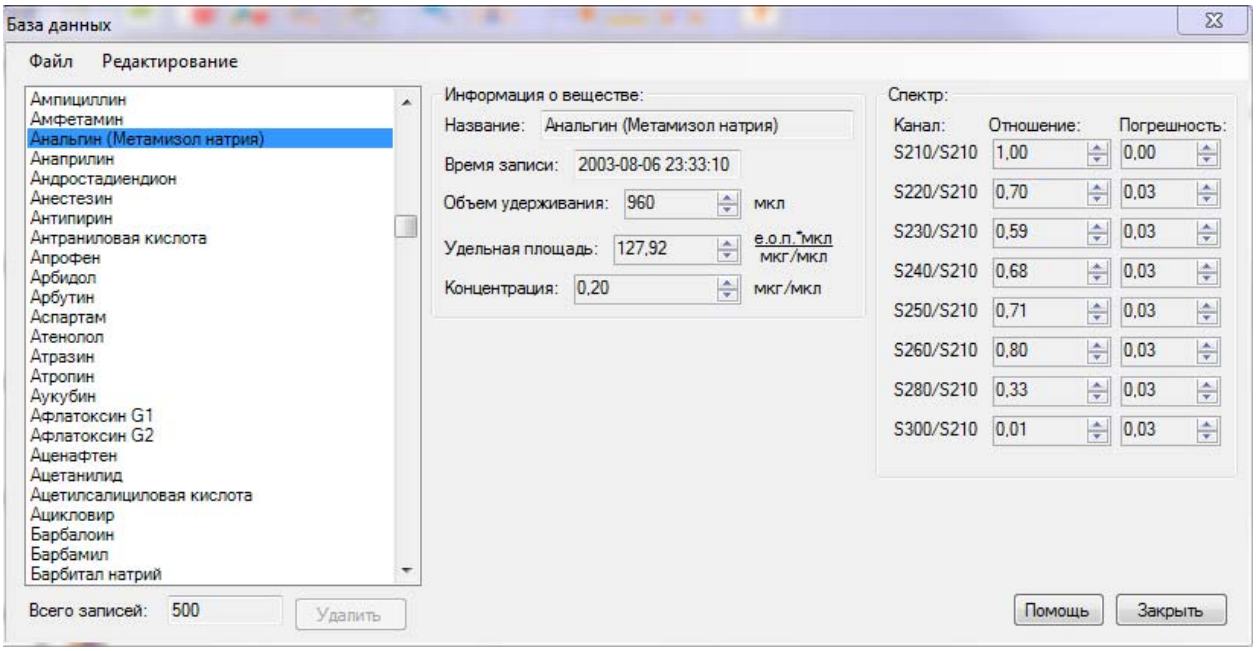


Рис. 11. Окно базы данных ВЭЖХ-УФ

Определение компонента производится следующим образом:

1. Определяется нормированный спектр и время удерживания рассматриваемого пика.
2. Из базы данных выбираются все записи, время удерживания которых попадает в заданное окно.
3. Из этих записей выбирается компоненты со спектрами, наиболее близкими к рассматриваемому.

Определение может проводиться как оператором в ручном режиме по каждому пику в отдельности (см. рис. 12), так и автоматически по всей хроматограмме (см. пример спектрального отчета на стр. 43). Есть возможность выводить не все подходящие записи из базы, а только т.н.

«наилучшего кандидата» — компонент, время удерживания и спектр которого наиболее близки к определяемому.

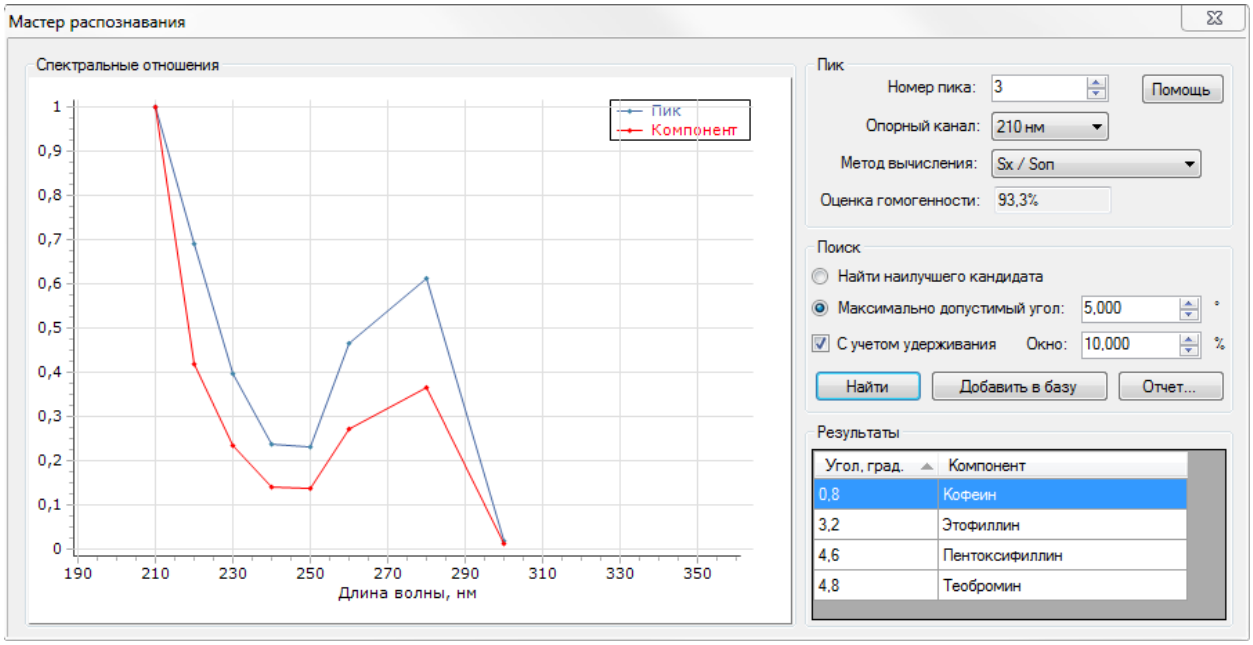


Рис. 12. Окно мастера определения компонента по базе данных.

8. СРАВНЕНИЕ ХРОМАТОГРАММ

На текущий момент нашей основной задачей является разработка метода качественного сравнения полученных наборов пиков для получения ответа на вопрос: "В чем именно заключается сходство и различие двух хроматограмм?". В качестве предварительного решения нами предлагается последовательное сравнение времен удерживания и спектральных векторов пиков.

8.1 Поиск одинаковых пиков

На первом шаге алгоритма обе хроматограммы разбиваются на пики, после чего строится таблица соответствия:

Таблица 1. Пример таблицы соответствия пиков.

	Пик 1	Пик 2	Пик 3
Пик A	(2,1 мин; 6,2 °)	(1,5 мин; 11,9°)	(2,0 мин; 0,1°)
Пик B	(0,4 мин; 0,7°)	(0,1 мин; 0,4°)	(3,1 мин; 8,0°)
Пик C	(0,3 мин; 0,2°)	(3,0 мин; 4,1°)	(0,2 мин; 0,2°)

Пики первой хроматограммы обозначаются как A, B, и C, пики второй хроматограммы — как 1, 2 и 3. В ячейках таблицы вписаны расстояния по времени между вершинами соответствующих пиков и углы их спектральных отклонений.

После этого, из таблицы удаляются ячейки, в которых разница по времени или спектру превышает заданные экспертом величины (см. табл. 2).

Таблица 2. Пример таблицы соответствия пиков после удаления невозможных кандидатов. В качестве пороговых значений выбраны 2,0 мин и 1,0°.

	Пик 1	Пик 2	Пик 3
Пик A			(2,0 мин; 0,1°)
Пик B	(0,4 мин; 0,7°)	(0,1 мин; 0,4°)	
Пик C	(0,3 мин; 0,2°)		(0,2 мин; 0,2°)

На последнем этапе выделяются наилучшие кандидаты. Для этого в таблице выбирается запись с минимальным углом отклонения (Пик A —

Пик 3 в примере на табл. 3), помечается как наилучший кандидат и вычеркивается, вместе со всей строкой и всем столбцом. Затем операция повторяется до тех пор, пока в таблице не останется записей.

Таблица 3. Пример таблицы соответствия пиков с найденными соответствиями.

	Пик 1	Пик 2	Пик 3
Пик A			(2,0 мин; 0,1°)
Пик B	(0,4 мин; 0,7°)	(0,1 мин; 0,4°)	
Пик C	(0,3 мин; 0,2°)		(0,2 мин; 0,2°)

Иными словами, если говорить в терминах теории графов, здесь реализуется жадный алгоритм поиска паросочетаний минимального веса, где в качестве вершин берутся пики, а в качестве весов ребер — углы между соответствующими этим пикам спектрами.

Таким образом строится словарь соответствий между пиками двух хроматограмм. В том случае, когда необходимо построить соответствие между несколькими хроматограммами (например, для выявления одинаковых пиков на наборе эталонных образцов) этот процесс выполняется последовательно для каждой предложенной хроматограммы.

Проиллюстрируем этот шаг на ещё одном простом примере. На рис. 13 изображены две хроматограммы, состоящие из трёх пиков:

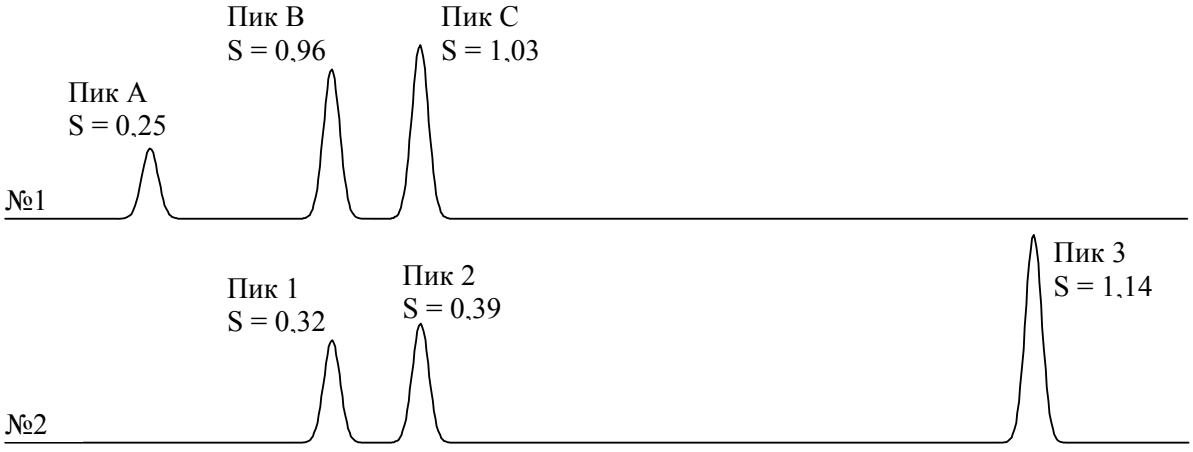


Рис. 13. Две хроматограммы с обозначенными номерами и площадями пиков.

Предположим, что спектры пиков B и 1 пиков C и 2 совпадают с достаточной точностью. Полученный словарь соответствий представлен в табл. 4:

Таблица 4. Словарь соответствий пиков для примера с рис. 13.

Хроматограмма №1	Хроматограмма №2
Пик A (S = 0,25)	—
Пик B (S = 0,96)	Пик 1 (S = 0,32)
Пик C (S = 1,03)	Пик 2 (S = 0,39)
—	Пик 3 (S = 1,14)

Таким образом, появляется возможность записать обе хроматограммы в виде векторов:

$$C_1=(0{,}25,\ 0{,}96,\ 1{,}03,\ 0{,}0) \text{ и } C_2=(0{,}0,\ 0{,}32,\ 0{,}39,\ 1{,}14),$$

где в качестве элементов используются значения площадей соответствующих пиков. Представление хроматограммы в подобном виде позволяет легко применять различные алгоритмы для сравнения и выделения кластеров, о которых будет сказано далее.

8.2. Выделение значимых признаков

После нахождения соответствий между пиками в одном наборе хроматограмм выбираются пики, которые можно считать значимыми признаками того или иного класса. Это необходимо для избавления от случайных примесей и сокращения размерности получаемого пространства. Мы предлагаем считать значимыми признаками пики, удовлетворяющие следующему условию: разность частот появления пика в различных источниках не менее 50% (см. скриншот на рис. 14). При этом размерность рассматриваемых векторов уменьшается с 300—400 до 20—30, т.е. как минимум на порядок. Эти «очищенные» вектора формируют так называемые «отпечатки пальцев», описывающие наиболее характерные признаки для определенного класса веществ.

	Среднее время, мин	Средний спектр	Совпадения (источник 1)	Совпадения (источник 2)	Признак
	23,349905	0,61842...	80,8%	00,0%	80,8
	12,976313	0,36457...	00,0%	61,5%	-61,5
	26,012059	0,29823...	03,8%	65,4%	-61,5
	14,761400	0,45497...	57,7%	00,0%	57,7
	08,469121	0,72249...	19,2%	73,1%	-53,8
	28,785929	0,15113...	00,0%	53,8%	-53,8
	01,177143	0,70272...	53,8%	00,0%	53,8
	22,749286	0,56250...	53,8%	00,0%	53,8
	02,471960	1,00398...	34,6%	84,6%	-50,0
	22,256250	0,59549...	57,7%	07,7%	50,0

Рис. 14. В первой строке таблицы описан пик, встречающийся в 80,8% хроматограмм из источника 1 (сыворотка крови доноров), и не встречающийся ни в одной из хроматограмм из источника 2 (сыворотка крови людей с онкологическими заболеваниями). Этот пик можно считать значимым признаком для первого источника.

8.3. Кластеризация и сравнение хроматограмм

В результате описанных выше действий мы получаем хроматограмму уже не в виде нескольких массивов пар «время — оптическая плотность» или набора времен удерживаний, площадей и спектров, а в виде вектора, элементами которого являются значения площадей соответствующих пиков.

Набрав достаточное количество хроматограмм сыворотки крови здоровых и больных людей, мы можем использовать их в качестве обучающих выборок для алгоритмов кластеризации. Для этого в разработанной системе применяются два основных алгоритма: алгоритм нечеткой кластеризации C-means [16] и метод ближайших соседей, в котором в качестве метрики для поиска эталонов используется функция конкурентного сходства FRiS [17]:

$$S(u,x \mid x') = \frac{\rho(u,x') - (\rho(u,x)}{\rho(u,x') + (\rho(u,x)} \qquad (12)$$

После создания обучающих выборок (сыворотка крови здоровых людей — сыворотка крови людей с онкологическими заболеваниями)

появляется возможность определить, к какому кластеру принадлежит рассматриваемая хроматограмма:

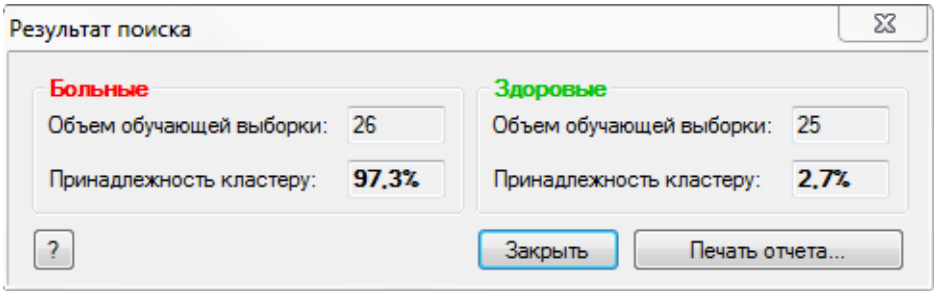


Рис. 15. Окно с результатом определения наличия заболевания.

9. ОЦЕНКА БЫСТРОДЕЙСТВИЯ

Было проведено несколько тестов быстродействия созданной системы [18]. Исследования проводились на компьютере со следующими характеристиками: Intel Core i3 M350 2,27 GHz, 3 GB RAM, Microsoft Windows 7 Professional 32bit. Язык реализации системы — C#, Microsoft .NET Framework 2.0.

Результаты, полученные при тестировании, представлены в табл. 5 и 6.

Таблица 5. Время, затрачиваемое на разбиение хроматограммы на пики.

Количество пиков на хроматограмме	Средний результат из 50 экспериментов
10	176,4 мсек
50	205,1 мсек
100	267,3 мсек

Таблица 6. Время, затрачиваемое на перевод обучающей выборки из 25 разделенных на пики хроматограмм в векторное представление.

Количество пиков на хроматограмме	Средний результат из 50 экспериментов
10	287,5 мсек
50	456,2 мсек
100	1353,3 мсек

10. ЗАКЛЮЧЕНИЕ

В результате работы была создана программная платформа, включающая в себя блок управления хроматографом и блок обработки получаемых хроматограмм (см. рис. 16). Процесс выделения информации из входных данных включает в себя фильтрацию сигнала детектора при помощи сглаживающих фильтров и вейвлет-преобразований, разбиение хроматографической кривой на пики и вычисление их характеристик. Дальнейшая обработка полученной информации заключается в обнаружении одинаковых пиков в разных хроматограммах путем метрической классификации их спектров с последующим переводом хроматограмм в векторное представление, понижением их размерности и применением к ним алгоритмов нечеткой кластеризации. Достоверность ответов системы зависит от задаваемых экспертом параметров алгоритмов выделения информации и качества и объема обучающих выборок.

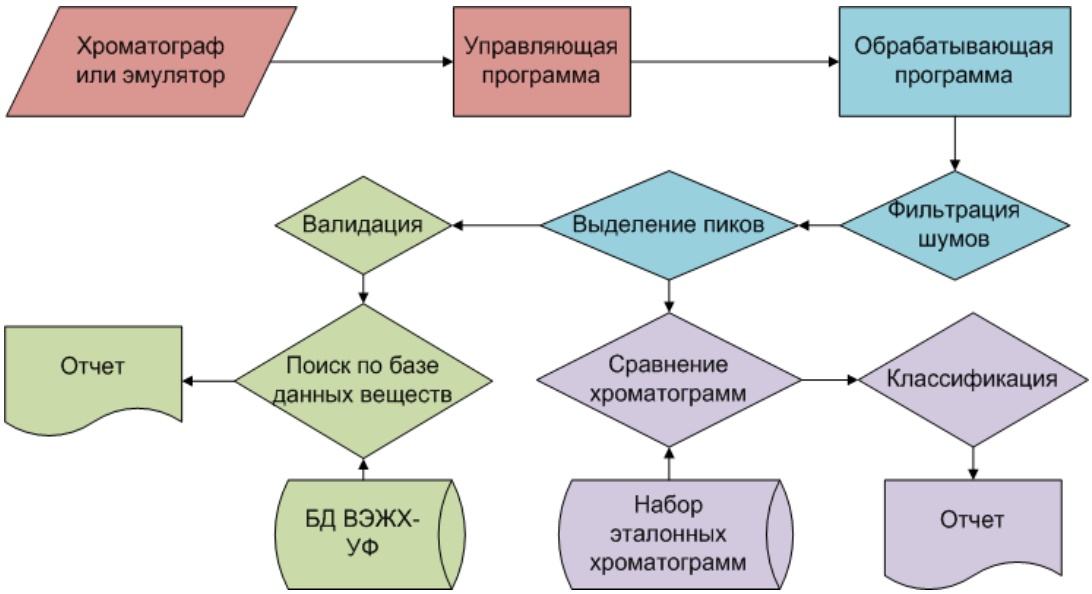


Рис. 16. Схема разработанной системы.

На текущий момент в мире существует достаточно много программных систем для обработки хроматограмм, реализующих те или иные алгоритмы сравнения и поиска образцов в базе данных, среди которых *ChemStation* от компании *Agilent* (США), *Empower* от компании *Waters* (США), *ChromQuest*

от компании *Thermo Scientific* (США) и десятки других. К сожалению, все они имеют ряд существенных недостатков:

- Идентификация веществ идет только по времени удерживания, что, с учетом возможного дрейфа пиков, дает значительную погрешность в результатах.
- В редких случаях исследуются спектральных отношения, но не для идентификации пика, а лишь для проверки его чистоты.
- Разделение пиков проводится лишь при помощи методов долин и перпендикуляров, что в некоторых случаях дает недостаточно точные результаты.
- Эти системы позволяют работать только с отдельными пиками, возможность оперировать наборами целых хроматограмм не предусмотрена.

Предложенное нами программное решение лишено большинства известных недостатков и позволяет проводить диагностику заболеваний, основываясь на результатах анализа образцов сыворотки крови, состав которых заранее неизвестен.

В данное время отдельные компоненты системы доступны в следующем виде:

- Управляющая программа «АльфаХром» (ЗАО Институт хроматографии «ЭкоНова»).
- Эмулятор «ВЭЖХ-УФ пептидов» включен в образовательный кластер «Молекулярная биология» РОСНАНО.
- Программа-тренажер «Жидкостный хроматограф» используется в НГУ и РХТУ им. Менделеева для обучения студентов-химиков.

Данная разработка является продолжением начатых ранее исследований [18]. Дальнейшая работа будет направлена на построение баз данных по реальным образцам сыворотки крови, что подразумевает тесное взаимодействие не только с экспертами из числа химиков-аналитиков, но и с представителями работников здравоохранения [18, 19].

11. СПИСОК ЛИТЕРАТУРЫ

1. Цвет М.С. О новой категории адсорбционных явлений и о применении их к биохимическому анализу / Труды Варшавского общества естествоиспытателей, 1903, Том XIV, Отделение биологии, Протокол №6, с. 1-20.

2. Raymond P. W. Scott. Liquid Chromatography Detectors / Library for Science, LLC, 2003.

3. Heyden Y.V. Extracting Information from Chromatographic Herbal Fingerprints / LCGC Europe, September 2008, pp. 438-443.

4. Л. Хубер, Применение диодно-матричного детектирования в ВЭЖХ (Москва «Мир», 1993).

5. N. Dyson, Chromatographic Integration Methods, 2nd ed. (Royal Society of Chemistry, Letchworth, UK, 1998).

6. J. Dolan, Integration Problems (LCGC North America, Volume 27, Number 10, October 2009).

7. МультиХром для Windows 9x & NT, версия 1.5x-E. Руководство пользователя. («АМПЕРСЕНД» 1993-2009, «ЭкоНова» 1997-2009).

8. Померанцев А.Л., Родионова О.Е. Хемометрика в аналитической химии / Электронный ресурс, http://www.chemometrics.ru (свободный доступ).

9. Zeng Z-D., Liang Y-Z., Xu C-J. Comparing chemical fingerprints of herbal medicines using modified window target-testing factor analysis / Anal. Bioanal. Chem., 2005, Vol. 381, pp. 913-924.

10. Hansen P.W. Pre-processing method minimizing the need for reference analyses / J.Chemom., 2001, Vol. 15, p. 123.

11. Померанцев А.Л. Методы нелинейного регрессионного анализа для моделирования кинетики химических и физических процессов / Дис. д-ра физ.-мат. наук, , Москва, ИХФ РАН, 2003.

12. Азарова И.Н., Барам Г.И., Гольдберг Е.Л. Предсказание объемов удерживания и УФ-спектров пептидов в обращенно-фазовой ВЭЖХ / Биоорганическая химия, 2006, т.32, №1, с.56-63.

13. Savitzky A., Golay M.J.E. Smoothing and differentiation of data by simplified least squares procedures / An. Chem., 1964.

14. Свидетельство № 38-03 об аттестации МВИ. Хроматографические и спектральные параметры УФ-поглощающих веществ. Методика выполнения измерений методом высокоэффективной жидкостной хроматографии.

15. Свидетельство № 67-06 об аттестации МВИ. Массовая концентрация УФ-поглощающих веществ. Методика выполнения измерений методом высокоэффективной жидкостной хроматографии.

16. Bezdek J.C. Pattern Recognition with Fuzzy Objective Function Algoritms / New York, Plenum Press, 1981.

17. Борисова И.А., Загоруйко Н.Г. Функции конкурентного сходства в задаче таксономии / Знания-Онтологии-Теории (ЗОНТ-07), 2007.

18. Барам Е.Г. Экспертная хроматографическая система для анализа образцов сложного состава. / Материалы 50-й юбилейной научной студенческой конференции «Студент и научно-технический прогресс» 13–19 апреля 2012, с. 99.

19. Барам Е.Г. Экспертная система определения заболевания по хроматограмме образца сыворотки крови. / Молодежный сборник ИСИ СО РАН, 2012 (в печати).

12. ПРИЛОЖЕНИЕ

ОТЧЕТ

Общая информация:

Файл: *E:\Chromatography\Кровь\Хемометрия\sorted\donors\157_16330080.chrw*
Время записи: *13 ноя 2010, Сб 16:33:48*
Метод: *"БД-2003" (изменен)*
Оператор: *Кожанова*

Проба:

Объем: *50 мкл*
Название:

Колонка:
Номер: *2739*
Длина: *75 мм*
Диаметр: *2 мм*
Адсорбент: *Пронтосил, Базы данных*
Размер частиц: *5 мкм*

Элюенты:

А: *[4М LiClO4 - 0.1М HClO4]:H2O= (5:95)*
Б: *MeCN ("Криохром", сорт 1)*
Максимальное давление: *2,7 МПа*
Скорость потока: *100 мкл/мин*
Температура: *40 °C*

График:

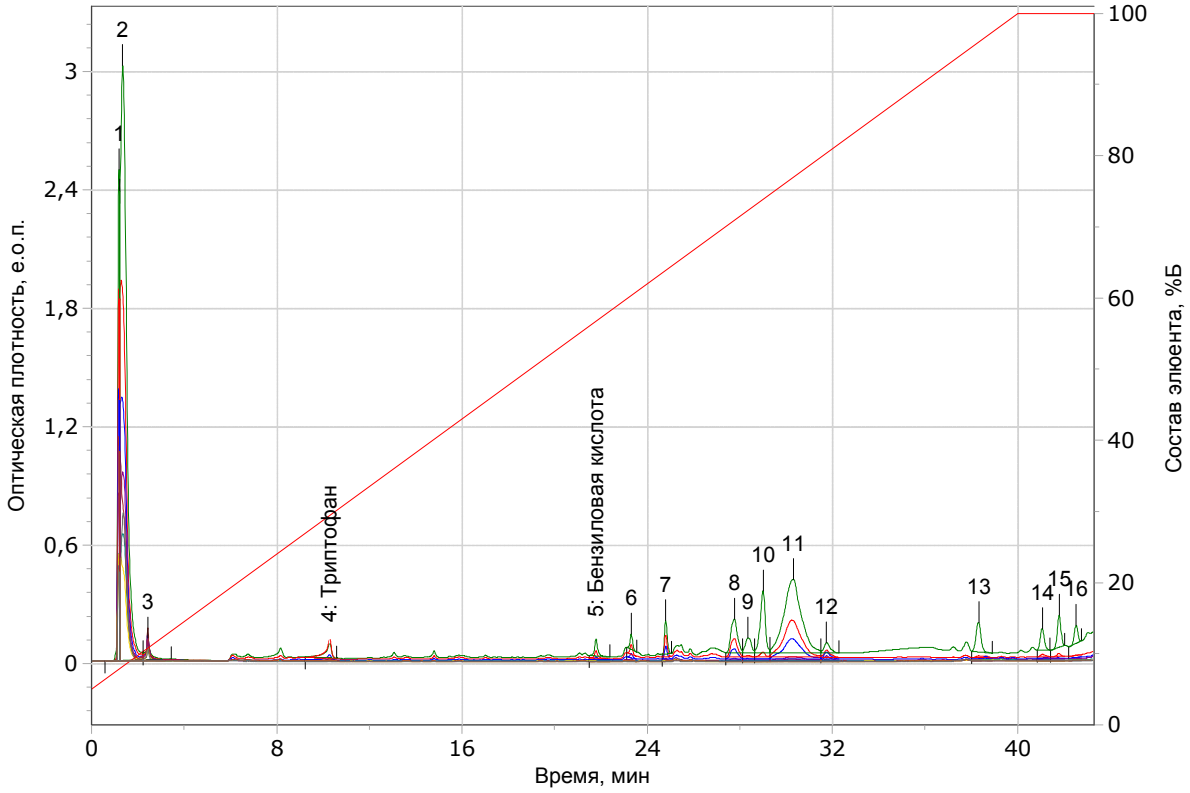


Таблица пиков (опорный канал 210 нм):

Условные обозначения:

Tr - время удерживания
h - высота пика
w - ширина пика при h/2
S - площадь пика
k' - фактор емкости
Rs - разрешение пиков n и n+1
N - эффективность колонки
ВЭТТ - высота, эквивалентная теоретической тарелке
A - асимметрия
C - концентрация

№	Tr	h	w	S	S	k'	Rs	N
	мин	е.о.п.	мин	мин*е.о.п.	%			т.т.
1	1,18	2,5	0,08	0,21133	11,109	-0,213	0,469	1120
2	1,34	3,03	0,31	0,99029	52,057	-0,109	2,962	105
3	2,42	0,12	0,13	0,02515	1,322	0,615	28,636	2073
4	10,28	0,09	0,2	0,02306	1,212	5,854	42,097	14981
5	21,79	0,11	0,12	0,01402	0,737	13,526	7,676	170938
6	23,31	0,14	0,11	0,00967	0,508	14,54	7,658	251877
7	24,8	0,21	0,12	0,02229	1,172	15,534	8,12	237404
8	27,77	0,21	0,31	0,05482	2,882	17,511	1,159	44420
9	28,36	0,12	0,29	0,02241	1,178	17,905	1,492	53350
10	29,01	0,36	0,23	0,07982	4,196	18,34	1,538	91777
11	30,31	0,42	0,77	0,32819	17,252	19,209	1,635	8515
12	31,74	0,1	0,26	0,01612	0,847	20,162	15,908	85740
13	38,31	0,2	0,23	0,03742	1,967	24,541	7,869	152576
14	41,06	0,17	0,18	0,02171	1,141	26,373	2,441	287259
15	41,79	0,23	0,17	0,02967	1,56	26,86	2,512	328064
16	42,52	0,18	0,17	0,01635	0,859	27,347		345208
Сумма				1,90232	100,0			

№	ВЭТТ	A	C	Имя
	мкм		мкг/мкл	
1	66,9	0,36		
2	713,5	3,05		
3	36,2	1,83		
4	5,0	0,25	1,928	Триптофан
5	0,4	1,26	2,151	Бензиловая кислота
6	0,3	0,95		
7	0,3	1,4		
8	1,7	1,13		
9	1,4	1,06		
10	0,8	0,73		
11	8,8	0,92		
12	0,9	1,53		
13	0,5	0,99		
14	0,3	1,16		
15	0,2	0,98		
16	0,2	1,02		
Сумма				

Параметры детектора:

Метод измерения: *однолучевой*
Рабочая кювета: *верхняя*
Регистрация данных: *после ввода пробы*
Уровень шума по каналам:

210нм	220нм	230нм	240нм	250нм	260нм	280нм	300нм
0,00011	0,00008	0,00008	0,00008	0,00010	0,00013	0,00016	0,00013

Параметры разметки:

Канал: *210 нм*
Начало: *0,0 мин*
Конец: *43,235 мин*
Минимальная ширина: *0,0 мин*
Минимальная высота: *0,05 е.о.п.*
Минимальная площадь: *0,0 мин*е.о.п.*
Базовая линия: *ортогональная*

Таблица компонентов:

№	Удерживание мин	Окно %	Концентрация мкг/мкл	Имя
1	10,31	5	1,928	Триптофан
2	21,76	5	2,151	Бензиловая кислота

№	Канал интегрирования нм
1	210
2	210

Спектральные отношения:

Опорный канал: *210нм*

Пик №	220нм	230нм	240нм	250нм	260нм	280нм	300нм
1	0,742	0,550	0,346	0,222	0,221	0,470	0,199
2	0,632	0,452	0,324	0,217	0,167	0,292	0,236
3	0,884	0,865	0,694	0,482	0,474	0,983	0,426
4	1,286	0,364	0,083	0,096	0,151	0,214	0,018
5	0,447	0,176	0,043	0,009	0,002	0,022	0,025
6	0,638	0,300	0,074	0,037	0,035	0,057	0,011
7	0,642	0,374	0,065	0,023	0,037	0,069	0,004
8	0,565	0,321	0,066	0,020	0,023	0,042	0,006
9	0,075	0,011	0,024	0,016	0,004	0,002	0,003
10	0,079	0,001	0,001	0,001	0,000	0,002	0,001
11	0,515	0,287	0,058	0,015	0,018	0,034	0,004
12	0,703	0,699	0,448	0,175	0,098	0,108	0,007
13	0,097	0,074	0,057	0,031	0,037	0,011	0,000
14	0,147	0,039	0,031	0,016	0,005	0,004	0,002
15	0,097	0,011	0,004	0,002	0,000	0,001	0,001
16	0,026	0,007	0,007	0,001	0,000	0,004	0,003

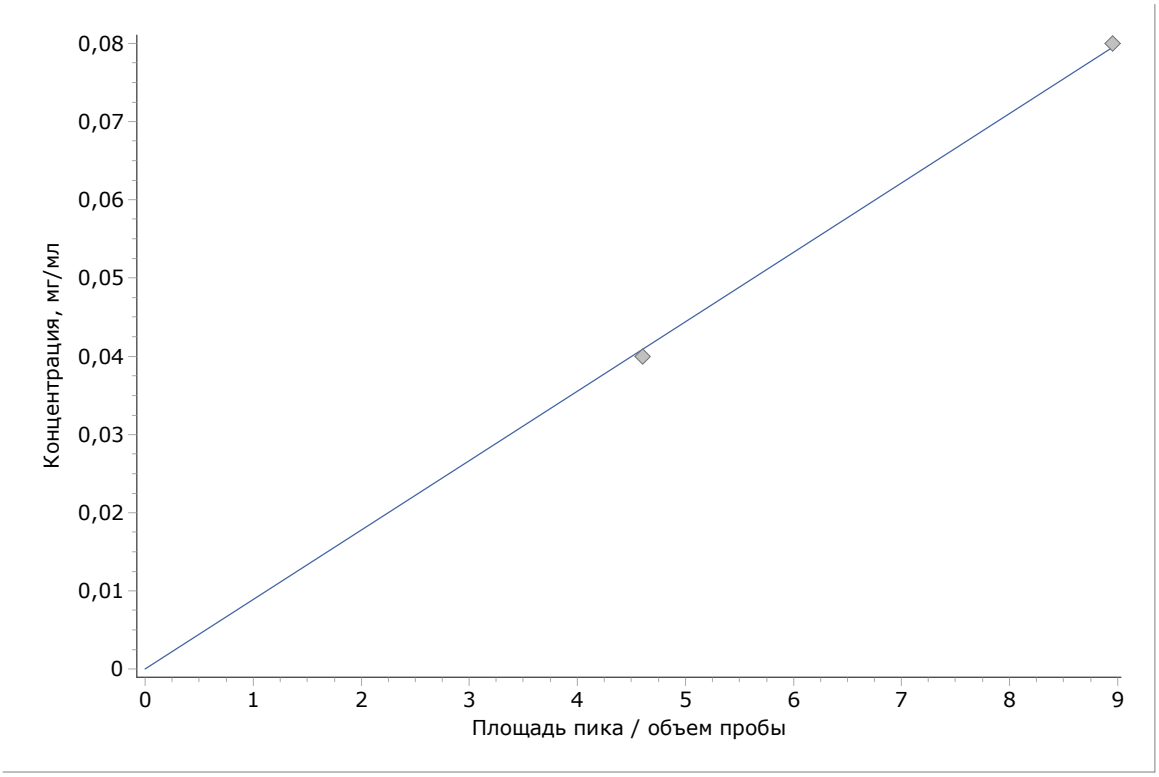
Результаты градуировки:

Компонент: *Кофеин*

Градуировочная зависимость: *C = 0,009 * S / V*

R^2: *1,0*

СКО: *1,6%*



Опорный канал: *210 нм*

Формула: *Линейная через 0*

Точка	Концентрация мг/мл	Площадь е.о.п.*мкл	Объем пробы мкл	Время выхода мин
1	0,04	9,2118	2	3,07
	C:\Alphachrom\Анализы\Кофеин\2012-06-20\20-07-04_001.chrx			
2	0,08	17,9069	2	3,04
	C:\Alphachrom\Анализы\Кофеин\2012-06-20\20-07-10_002.chrx			

СПЕКТРАЛЬНЫЙ ОТЧЕТ

Дата создания отчета: *6 апр 2012, Пт 19:19:44*
Файл: *E:\Chromatography\Кровь\Хемометрия\sorted\donors\157_16330080.chrw*
Дата записи файла: *13 ноя 2010, Сб 16:33:48*
Метод анализа: *"БД-2003".mtd*
Метод обработки: *БД-2003.mtdw (изменен)*
Продолжительность анализа: *4330 мкл (43,3 мин)*
Оператор: *Кожанова*
Номер анализа: *7*
Проба: *50 мкл, пробирка №8*
Колонка: *D = 2 мм, L = 75 мм, dp = 5 мкм, "Пронтосил, Базы данных", №2739*
Элюент А: *[4М LiClO4 - 0.1М HClO4]:H2O= (5:95)*
Элюент Б: *MeCN ("Криохром", сорт 1)*
Максимальное давление: *2,7 МПа*
Скорость потока: *100 мкл/мин*
Температура: *40 °C*

Параметры распознавания по спектрам
Доверительное окно объема удерживания: *10%*
Метод вычисления спектров: *Sx / Son*
Число эталонных спектров в базе: *500*

****
Номер пика: *4*
Удерживание: *1028 мкл (10,28 мин)*

Идентификация по спектру положительная:

Название	Объем	Угол	Концентрация
Триптофан	1029 мкл	0,3°	0,0 мкг/мкл

Спектральные отношения (относительно канала 210нм):

Название	220нм	230нм	240нм	250нм	260нм	280нм	300нм
Пик 4	1,301	0,365	0,084	0,093	0,154	0,217	0,018
Триптофан	1,354	0,378	0,086	0,093	0,158	0,230	0,020

****
Номер пика: *5*
Удерживание: *2179 мкл (21,79 мин)*

Идентификация по спектру положительная:

Название	Объем	Угол	Концентрация
Фенадон	2314 мкл	3,0°	0,001 мкг/мкл

Спектральные отношения (относительно канала 210нм):

Название	220нм	230нм	240нм	250нм	260нм	280нм	300нм
Пик 5	0,443	0,170	0,035	0,007	0,005	0,012	0,005
Фенадон	0,471	0,174	0,042	0,018	0,021	0,017	0,021