***Доклады Международной конференции "Математическая биология и биоинформатика"**. Под ред. В.Д. Лахно. Том 7. Пущино: ИМПБ РАН, 2018. Статья № e36. doi: 10.17537/icmbb18.50*

Метод оптимальных разбиений для оценки влияния степени оксигенации гемоглобина на фактор роста эндотелия сосудов

Сенько О.В.[1], Кодрян М.С.[2], Кузнецова А.В.[3], Клименко Л.Л.[4]

[1]*Федеральный исследовательский центр «Информатика и управление» РАН*
[2]*Московский государственный университет им. М.В. Ломоносова*
[3]*Институт биохимической физики им. Н.М. Эмануэля*
[4]*Институт химической физики им. Н.Н. Семёнова*

senkoov@mail.ru, max-kodr@rambler.ru, azfor@yandex.ru, klimenkoll@mail.ru

Целью работы является исследование связи фактора роста эндотелия сосудов VEGF в сыворотке крови с гипоксией в группах пациентов, страдающих тяжёлыми неврологическими заболеваниями. Имеющиеся в литературе данные свидетельствуют об активации синтеза VEGF при гипоксии. Вместе с тем стандартный корреляционный анализ не позволил достоверно выявить по клиническим данным наличие связи между уровнем VEGF и параметрами оксиметрии, объективно характеризующими снабжение организма кислородом. В статье представлена методика, позволившая статистически достоверно доказать существование указанной связи, сводящейся к увеличению корреляции между VEGF и комплементом C4, а также между VEGF и белками S100 при снижении уровня оксигенации ниже некоторого порогового значения. Методика основывается на построении оптимальных разбиений интервалов значений показателей оксиметрии. Использовалась оригинальная методика статистической верификации, основанная на использовании перестановочных тестов, которая позволила совместно учесть отдельные эффекты по группе показателей оксиметрии, а также провести коррекцию значимости с целью учёта множественного тестирования.

Ключевые слова: *фактор роста эндотелия сосудов, гипоксия, верификация, перестановочный тест, множественное тестирование*

Optimal Valid Partitioning Method in Task of Evaluating Effect of Hemoglobin Oxygenation Levels Vessel Endothelial Growth Factor

Senko O.V.[1], Kodryan M.S.[2], Kuznetsova A.V.[3], Klimenko L.L.[4]

[1]*FRC «Computer Science and Control» of RAS*
[2]*Lomonosov Moscow States University*
[3]*Emanuel Institute of Biochemical Physics of RAS*
[4]*Semenov Institute of Chemical Physics of RAS*

The work is aimed to study relationship between vessel endothelial growth factor (VEGF) serum levels and hypoxia in patients with severe neurological disorders. Overviewed literature sources indicate activation of VEGF synthesis during hypoxia. But standard correlation analysis does not reveal a statistically significant association between VEGF levels and pulse oxymetry parameters objectively assessing oxygen supply. Due to a proposed in the paper technique it became possible to reveal valid relationship between VEGF levels and oxygenation. This relationship can be described as increase of correlation between VEGF and immunological complement C4 when oxygenation level is below a certain threshold. The method is based on optimal partitioning of pulse oxymetry parameters ranges. The original technique based on permutation test was used to asses validity of joint effect by group of oxymetry parameters. This technique was applied for validity correction aimed to take into account multiple testing.

Key words: vessel endothelial growth factor, hypoxia, verification, permutation test, multiple testinge

1. Введение

Ключевым фактором, регулирующим процесс ангиогенеза, признан фактора роста эндотелия сосудов (vascular endothelial growth factor – VEGF) [1, 2]. В норме VEGF содержится в тканях в незначительном количестве (10–246 пг/мл), но экспрессия его гена значительно активируется при гипоксии, доходя в нашем исследовании до крайне высокого значения 3176 пг/мл. Тканевая гипоксия вызывает активацию генов семейства HIF-1 или гипоксия-индуцибельного фактора (hypoxia-inducible factor) [2]. Активация гена HIF-1 происходит в физиологически важных местах регуляции кислородных путей, обеспечивая быстрые и адекватные ответы на гипоксический стресс, в первую очередь – ответ генов, регулирующих процесс ангиогенеза и способствующих образованию VEGF, продуцируемого различными типами клеток – макрофагами, фибробластами, лимфоцитами, полиморфноядерными клетками и т.д. [2]. Несмотря на сложившееся понимание, что VEGF является важнейшим инструментом физиологической регуляции, в настоящее время существует ограниченное число работ, в которых влияние гипоксии и других биологических факторов на активацию синтеза VEGF оценивается по клиническим данным, что может быть связано со сложным и многоуровневым механизмом регуляции, делающих выявление статистически достоверных эффектов по отдельным переменным затруднительным. Например, стандартный корреляционный анализ не позволил достоверно выявить по анализируемым в настоящей работе клиническим данным наличие связи между уровнем VEGF и показателями оксиметрии, объективно характеризующими снабжение организма кислородом. Однако, как это показывается далее, успех при поиске закономерностей, связанных с регуляцией уровня VEGF, может быть достигнут при учёте взаимодействия различных показателей,

Поиск таких закономерностей затрудняет сложный и существенно нелинейный характер взаимодействия.

В этих условиях исследование процесса регуляции синтеза VEGF как реакцию на гипоксию, целесообразно проводить через поиск математических моделей связывающих уровень VEGF с показателями пальцевой оксиметрии в сочетании некоторым дополнительным фактором, принадлежащим набору разнообразных клиничеких, биохимических или инструментальных показателей, присутствующих в базе данных. При этом модели строятся в соответствии с возникающими предположениями о характере зависимостей. Важнейшей составляющей исследования является верификация выявленных эмпирических закономерностей, которая должна включать невозможность простого объяснения последних только одномерными эффектами. Верификация закономерностей осложняется также несостоятельностью гипотез о нормальности распределений для большинства биомедицинских показателей. Для верификации моделей целесообразно использовать перестановочные тесты, не требующие априорных предположений о распределениях и являющиеся универсальным инструментом статистической верификации [4].

2. Материалы и методы

2.1. Данные

Статистические исследования проводились на основе базы данных, содержащей значения клинических, лабораторных и инструментальных показателей для 88 пациентов с возрастом от 33 до 88 лет, страдающих неврологическими заболеваниями: острое нарушение мозгового кровообращения (ишемический инсульт) и транзиторная ишемическая атака.

База данных содержит значения 146 показателей, включая содержание VEGF в сыворотке крови (пг/мл); 7 показателей оксиметрии, включая парциальное давление кислорода в артериальной крови pO2 (мм. рт. ст.), парциальное давление углекислого газа в крови pCO2 (мм. рт. ст.), индекс сатурации sO2 (%), количественные значения биохимических показателей, показатели медленной электрической активности коры головного мозга – уровень постоянного потенциала, (УПП, мВ), концентрация микроэлементов в сыворотке крови (мкг/г).

Исследование проведено в клинической больнице № 123 ФМБА России.

Определение концентрации биохимических, гематологических и иммунологических показателей выполнялось в клинико-диагностической лаборатории на автоматическом биохимическом анализаторе «RX Imola» фирмы «Randox» (Великобритания), автоматическом биохимическом анализаторе «Сапфир-400», автоматическом гематологическом анализаторе «Medonic MC-15», «МЕК 7222», автоматическом иммуноферментном анализаторе «Лазурит» фирмы «Вектор-бест» (Россия), с использованием реагентов фирмы «Randox» (Великобритания), «CORMAY» (Польша), фирмы «Юнимед» (Россия), фирмы «Вектор-бест» (Россия), соответственно. Пробы крови брали свободным истечением из локтевой вены утром натощак через 12–14 часов после приема пищи.

Для определения концентрации нейроспецифических белков VEGF (пг/мл) и S100 (нг/л), а также белков сывороточного комплемента C3 и C4 (г/л) с помощью прибора «Лазурит» использовали метод иммуноферментного анализа ELISA (enzyme linked immunosorbent assay). Далее содержания VEGF будет обозначаться Y. Показатели оксиметрии будут обозначаться $X_1, X_2,\ldots, X_g$, где $g = 7$. Остальные показатели,

вошедшие в базу и являющиеся потенциальными дополнительными факторами, будут обозначаться $Z_1,\ldots, Z_n$, где n в нашем случае рано 128.

Анализируемая выборка может быть представлена в виде

$$\tilde{S}_o = \{(y_1, \mathbf{x}_1, \mathbf{z}_1),\ldots,(y_m, \mathbf{x}_m, \mathbf{z}_m)\},$$

где y_j – значение переменной Y, $\mathbf{x}_j$ – вектор значений переменных $X_1,\ldots, X_g$, $\mathbf{z}_j$ – вектор значений переменных $Z_1,\ldots, Z_n$

Стандартный корреляционный анализ не позволил выявить сколь-либо достоверной корреляции между VEGF и объективно характеризующими гипоксию показателями оксиметрии. Результаты использования метода оптимальных достоверных разбиений [3] и визуальный анализ диаграмм рассеяний позволил предположить, что такая связь может быть описана с помощью условно-линейных моделей.

2.2. Метод достоверных условно-линейных закономерностей. Оптимизируемый функционал

Для поиска закономерностей использовался метод достоверных условно-линейных закономерностей (МДУЛЗ), являющийся модификацией метода оптимальных достоверных разбиений. МДУЛЗ позволяет оценить существовании достоверной совместной связи переменных X_i, Y и Z_j. В основе МДУЛЗ лежит предположение о значительном увеличении модуля коэффициента корреляции между Y и Z в группе пациентов, для которых X находится с одной стороны от некоторого порога δ, по сравнению с модулем коэффициента корреляции в группе с X, находящемся с другой стороны порога. Соответствие такого предположения данным может быть охарактеризовано функционалом

$$\Phi(\tilde{S}_o, \delta) = \frac{|| \rho_l(Y,Z)| - |\rho_r(Y,Z)|| m_l m_r}{\sqrt{[1-\rho^2_{\max}(Y,Z)]}},$$

где $\rho_l(Y, Z)$ коэффициент корреляции между Y и Z в группе со значениями X слева от порога δ,

$\rho_r(Y, Z)$ – коэффициент корреляции между Y и Z в группе со значениями X справа от порога δ,

m_l – число пациентов в группе со значениями X слева от порога δ,

m_r – число пациентов в группе со значениями X слева от порога δ,

$\rho_{\max}(Y, Z)$ максимальное значение из $\rho_l(Y, Z)$ и $\rho_r(Y, Z)$.

Оптимальное значение порога δ ищется через максимизацию функционала $\Phi(\tilde{S}_o)$.

Следует отметить наличие в данных наблюдений, значительно отклоняющихся от основных закономерностей, существующих в данных. Поэтому вместо стандартных коэффициентов корреляции использовались их робастные аналоги. По группам наблюдений слева и справа от порогового значения δ с использованием метода наименьших квадратов строились одномерные регрессии переменной Y по переменной Z и вычислялись среднеквадратичные отклонения σ реальных значений Y от рассчитанных по регрессионной формуле. В каждой из групп отбрасывались наблюдения, для которых величина отклонения превосходила 3σ. Коэффициенты корреляции, рассчитанные по полученным группам считались робастными.

В связи с тем, что при малом размере одной из групп слева или справа от порога резко возрастает вероятность случайного повышения модуля коэффициента корреляции до величин близких к 1, интервал поиска δ ограничивается условием $m_l, m_l > 25$.

2.3. Метод достоверных условно-линейных закономерностей. Верификация

Как уже говорилось ранее, верификация производится с помощью перестановочных тестов. При этом верификация отдельной закономерности для фиксированной тройки (X_i, Y, Z_j) производится с помощью стандартного варианта теста. Осуществить верификацию одновременно по всевозможным тройкам позволяет следующая схема I.

А) С помощью датчика случайных генерируется множество $\{f_k \mid k = 1,\ldots, N\}$, состоящее из случайных перестановок чисел из $\{1,\ldots, m\}$.

Б) По каждой перестановке f_k генерируем из $\tilde{S}_o$ случайную выборку $\tilde{S}^r_k = \{(y_{f_k(1)}, \mathbf{x}_1, \mathbf{z}_1),\ldots$ $\ldots, (y_{f_k(m)}, \mathbf{x}_m, \mathbf{z}_m)\}$.

В) По исходной выборке $\tilde{S}_o$ и по каждой выборке $\tilde{S}^r_k$ с номером $k \in \{1,\ldots, N\}$ для каждой тройки (X_i, Y, Z_j) вычисляем оптимальное пороговое значение δ_o и соответствующее ему максимальное значение $\Phi_o(\tilde{S}^r_k) = \Phi(\tilde{S}^r_k, \delta_o)$.

Г) Для каждой тройки (X_i, Y, Z_j) находятся характеристики значимость соответствующей закономерности, включая p-значение, вычисляемое по формуле:

$$p = \frac{|\{\tilde{S}^r_k \mid \Phi_o(\tilde{S}^r_k) \geq \Phi_o(\tilde{S}_o), k = 1,\ldots, N\}|}{N},$$

и h-значение, вычисляемое по формуле:

$$h = \frac{\Phi_o(\tilde{S}_o)}{\max_{k=1,\ldots,N} \Phi_o(\tilde{S}^r_k)}.$$

Характеристика h-значение предназначена для сравнения значимости двух закономерностей, для которых p-значения равны 0.

Закономерности, связанные с увеличением корреляции между Y и дополнительным фактором Z при наличии гипоксии, должны проявляться для различных параметров оксиметрии, имеющих сходный биологический смысл. Разумно предположить снижение вероятности случайного появления конфигураций данных, формально

соответствующих значимым закономерностям сразу для нескольких параметров при одном и том же дополнительном факторе *Z*. Поэтому разумно использовать оценку достоверности по всей группе параметров оксиметрии. Для оценки группового эффекта использовалась следующая схема II.

А) С помощью датчика случайных генерируется множество $\{f_k \mid k = 1,...,2N\}$, состоящее из случайных перестановок чисел из $\{1,\ldots, m\}$.

Б) По каждой перестановке f_k генерируем из $\tilde{S}_o$ случайную выборку $\tilde{S}_k^r = \{(y_{f_k(1)}, \mathbf{x}_1, \mathbf{z}_1),\ldots$ $\ldots,(y_{f_k(m)}, \mathbf{x}_m, \mathbf{z}_m)\}$.

В) По исходной выборке $\tilde{S}_o$ и по каждой выборке $\tilde{S}_k^r$ с номером $k \le N$ вычисляем для каждой тройки (X_i, Y, Z_j) соответствующие *p*-значения и *h*-значения. Вычисления проводятся помощью схемы, аналогичной схеме I, но с использованием в качестве случайных выборок все выборки с номерами $k > N$.

Г) По исходной выборке $\tilde{S}_o$ и по каждой выборке $\tilde{S}_k^r$ с номером $k \le N$ для каждого дополнительного фактора Z_j вычисляем интегральный показатель значимости. В качестве таковых могут выступать сумма P_j^s или медиана P_j^{med} по *p*-значениям для всех троек с фиксированным Z_j.Также могут использоваться сумма H_j^s или медиана H_j^{med} по *h*-значениям для всех троек с фиксированным Z_j.

Д) Находятся характеризующие значимость дополнительного фактора *p*-значение и *h*-значение, которые при использовании в качестве интегрального фактора медианы H_j^m будут вычисляться по формулам:

$$p_j^m(\tilde{S}_o) = \frac{|\{\tilde{S}_k^r \mid H_j^m(\tilde{S}_k^r) \ge H_j^m(\tilde{S}_o), k = 1,\ldots,N\}|}{N}, \quad (1)$$

$$h_j^m(\tilde{S}_o) = \frac{H_j^m(\tilde{S}_o)}{\max_{k=1,\ldots,N} H_j^m(\tilde{S}_k^r)}. \quad (2)$$

Важную роль при оценивании значимости в условиях высокой размерности имеет эффект множественного тестирования (ЭМТ), связанный с возрастанием вероятности чисто случайного появления конфигураций данных, формально соответствующих статистически достоверным закономерностям. На самом деле какая-либо связь между соответствующими переменными отсутствует. В нашем случае ЭМТ сводится к случайному появлению среди потенциальных дополнительных факторов показателя Z_j с низким *p*-значением согласно формуле (1) или, соответственно, высоким *h*-значением согласно формуле (2). При этом какая-либо связь между Y и сочетанием Z_j с показателями оксиметрии отсутствует. Для того чтобы надёжно убедиться в реальном существовании зависимости, необходимо проводить дополнительную коррекцию на множественное тестирование.

Для осуществления такой коррекции была предложена схема, основанная на генерации трёх групп случайных выборок, получаемых с помощью перестановок позиций Y относительно фиксированных позиций векторов **x** и **z**. То есть помощью датчика случайных генерируется множество $\{f_k \mid k = 1,...,3N\}$, состоящее из случайных перестановок чисел из $\{1,\ldots, m\}$.

По каждой перестановке f_k генерируем из $\tilde{S}_o$ случайную выборку $\tilde{S}_k^r = \{(y_{f_k(1)}, \mathbf{x}_1, \mathbf{z}_1),\ldots$ $\ldots,(y_{f_k(m)}, \mathbf{x}_m, \mathbf{z}_m)\}$.

Интегральные значимость для каждого дополнительного фактора оценивается по $\tilde{S}_o$ и каждой из сгенерированных случайных выборок с номерами $N < k \le 2N$ по формулам (1) и (2). Однако расчёты проводятся не по выборкам с номерами $k \le N$, как это делается в схеме II, а по выборкам с номерами $N < k \le 2N$. Совместная связь фактора Z_j и параметров оксиметрии с Y считается значимой на уровне α, если доля выборок с номерами $k \le N$, для которых выполняется следующее условие спонтанной значимости, не превышает α. Условие спонтанной значимости:

существует такой фактор Z_{j^*}, для которого выполняются условия $p_{j^*}^m(\tilde{S}_k^r) \le p_j^m(\tilde{S}_o)$ или условие $h_{j^*}^m(\tilde{S}_k^r) \le h_j^m(\tilde{S}_o)$.

Перечисленные выше методы позволяют не только учесть групповой эффект, но также провести коррекцию на множественное тестирование. Однако, для исчерпывающей оценки достоверности выявленной связи VEGF с оксигенацией в сочетании с дополнительным фактором Z_{j^*} не достаточно проверки одной только нулевой гипотезы о независимости Y от совместного распределения переменных $X_1,\ldots,X_g$ и Z_{j^*}. Действительно, эта одна лишь эта проверка не исключает возможность случайного возникновения наблюдаемой конфигурации данных вследствие реально существующей связи между Y и Z_{j^*}. Исчерпывающей оценке достоверности соответствует дополнительной проверке ещё двух нулевых гипотез: независимости совместного распределения переменных Y и Z_{j^*} от переменных $X_1,\ldots, X_g$, независимости совместного распределения переменных Y и $X_1,\ldots, X_g$ от переменной Z_{j^*}. Проверка этих двух нулевых гипотез проводится совершенно аналогично.

3. Результаты

Были выявлены достоверные закономерности, связанные с увеличением корреляции между содержанием VEGF в сыворотке крови с двумя дополнительными факторам при наличии гипоксии. Такими дополнительными факторами являлся уровень комплемента C4 и содержание протеинов S100. Гипоксия фиксировалась по изменению параметров оксиметрии. При проверке нулевой гипотезы о независимости VEGF от сочетания параметров оксиметрии и уровня C4 достоверность выявленного эффекта для C4 с учётом группового эффекта определялась *p*-значением, меньшим 0.0001, и *h*-значением, равным 1.056. После коррекции на множественное тестирование значимость была оценена на уровне $p < 0.0082$. Скорректированная значимость для S100 была оценена на уровне $p < 0.012$.

Изменение корреляции между содержанием VEGF уровнем комплемента C4 иллюстрируется рисунками 1 и 2. На рисунке 1 показана связь между C4 и VEGF справа от границы для индекса сатурации sO2, равной 39.25%.

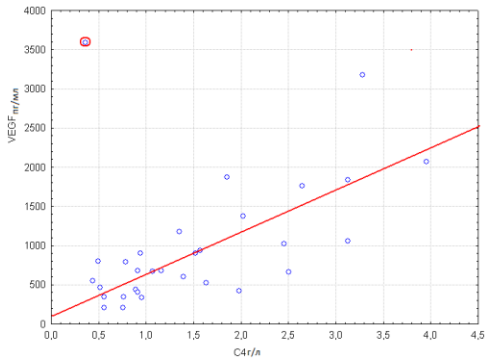

Рис. 1. Связь между C4 и VEGF при sO2 < 39.25 %.

Коэффициент корреляции в группе из 31 случая при sO2 < 39.25 % составил 0.47. После исключения выпадающего наблюдения, обведённого на рисунке кружком коэффициент корреляции вырос до 0.76

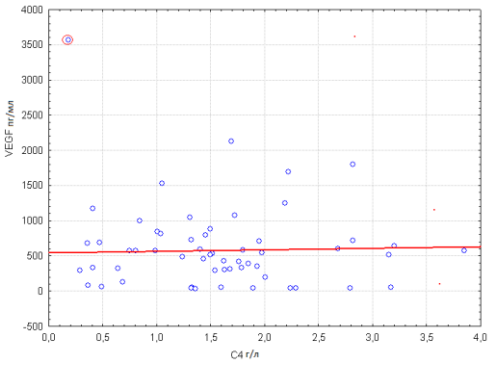

Рис. 2. Связь между C4 и VEGF при sO2 > 39.25 %.

Коэффициент корреляции в группе из 57 случаев при sO2 > 39.25 % составил –0.11. После исключения выпадающего наблюдения, обведённого на рисунке кружком коэффициент корреляции вырос до 0.05.

4. Заключение

Разработан новый метод поиска и верификации закономерностей, связанных с возрастанием корреляции между двумя показателями при превышении или при недостижении некоторого порога третьим показателем. При использовании разработанной технологии был выявлен статистически достоверный эффект, характеризующий связь уровня VEGF с гипоксией. Эффект состоит в повышении корреляции между VEGF и C4, а также между VEGF и S100, при гипоксии.

5. Благодарности

Работа была выполнена благодаря поддержке РФФИ (грант 17-07-01362).

6. Список литературы

1. Adair J.P. *Montani Angiogenesis*. Morgan & Claypool Life Science, 2010.
2. Захарова Н., Воскресенская О., Тарасова Ю. Ангиогенез и фактор роста эндотелия сосудов при цереброваскулярной патологии. *Врач. Научно-практический журнал*. 2014. № 10.
3. Кузнецова А.В., Костомарова И.В., Сенько О.В. Логико-статистический анализ связи клинико-лабораторных показателей с возникновением нарушения мозгового кровообращения у пациентов пожилого возраста с хронической ишемией головного мозга. *Матем. биология и биоинформ*. 2013. Т. 8. № 1. С. 182–224.
4. Pesarin F., Salmaso L. *Permutation Tests for Complex Data. Theory, Applications and Software*. John Wiley and Sons, Ltd, 2010.