Оптимизация факторной области модели регуляции экспрессии генов

С. Н. Истомина

***Аннотация* — Работа посвящена изучению множества изменяемых параметров математической модели регуляции экспрессии генов с целью выделить из них те, что служат факторами, влияющими, на результат счета. Задача состояла также в определении оптимальных условий для улучшения характеристики качества регуляции в результате счета. Для оптимизации эксперимента проводилось ортогональное и симплексное планирование. Для повышения надежности и достоверности выводов использован многомерный корреляционный анализ и уравнения регрессии со взаимодействиями факторов второго порядка, адекватно описывающие стационарную область.**

***Ключевые слова* — математическая модель, экспрессия, регуляция, гены, оптимальное планирование, ортогональность, симплекс, корреляционный анализ, регрессионный анализ.**

I. Введение

В настоящее время строго математически сформулированные модели в геномике используются и имеют значение. Методы планирования эксперимента были нами применены для изучения области параметров одной из таких моделей – процесса аттенюаторной регуляции экспрессии генов у бактерий [1], [2]. Массовый счет на основе такой модели позволяет сравнивать компьютерные результаты с известными экспериментальными и биоинформатическими данными и, тем самым, с одной стороны, оценить модель, а с другой стороны, получить новые предсказания аттенюаторных регуляций в отдельных бактериях. Модель также служит средством для оценки эффективности предсказанной регуляции.

Проблема состоит в том, что большинство параметров модели экспериментально оцениваются с низкой точностью, т.е. известны интервально; поэтому представляет интерес использовать регрессионный анализ, чтобы исследовать область значений параметров, а также определить, какие из них наиболее существенно влияют на изучаемую регуляцию, найти взаимосвязь параметров модели между собой (взаимодействие двух и более параметров), определить характер влияния одновременного изменения нескольких параметров модели на качество регуляции. Для этого необходимо обеспечить надежность уравнений регрессии и их коэффициентов, что может быть сделано при использовании планирования условий счета на модели с использованием критериев оптимальности.

Математическая модель при описании этой регуляции использует более 30 параметров. На основе экспертной оценки из них для исследования были выделены 10 параметров, вклад которых в процесс регуляции наиболее значителен. Далее эти параметры называются факторами и перечисляются вместе с обозначениями:

1) поправка для энергии связи микросостояния (x_1),
2) характеристика упорядоченности спиралей при выводе их списка (x_2),
3) константа скорости перехода полимеразы на следующий нуклеотид (x_3),
4) константа скорости перехода рибосомы на следующий кодон (x_4),
5) константа скорости срыва полимеразы (x_5),
6) константа замыкания (x_6),
7) доля урацилов в полимеразе (x_7),
8) длина полимеразы от места выхода цепи РНК до точки транскрипции (x_8),
9) длина рибосомы от ее Р-участка (центра) до ее 3'-края (x_9),
10) характеристика 5'-края полимеразы при начале моделирования (x_{10}).

II. Результаты ортогонального планирования

В результате модельного счета на выходе получается набор вероятностей для значения концентрации c аминоацил-тРНК синтетазы, которая меняется с шагом 0.05 в интервале от 0 до 1, то есть качество регуляции характеризует кривая изменения вероятностей при увеличении концентрации $p(c)$. Чтобы решать поставленные задачи для оценки качества регуляции, нужен числовой параметр Y с определенными свойствами. В качестве такого параметра регуляции Y приняли число равное приращению вероятности $p(c)$ на участке монотонного возрастания с учетом некоторых условий. Значение параметра равное нулю означает отсутствие регуляции, а чем больше значение, тем качество регуляции лучше.

Воспроизводимость результатов счета, оцениваемых по такому параметру, оказалась приемлемой для их статистического анализа.

Если каждый из 10 факторов варьировать хотя бы на двух уровнях, то возникают 1024 опыта с различными условиями, что приводит к трудностям при их реализации и построении адекватной регрессии. Поэтому для пошаговой стратегии оптимального планирования использованы ортогональные реплики от полного плана с двумя уровнями [3].

Статья получена 2 декабря 2014.

С. Н. Истомина, Российский государственный технологический университет имени К. Э. Циолковского, Москва, Россия (e-mail: istominasn@mail.ru).

Минимальный объем такой реплики 32 опыта, которых достаточно, чтобы независимо оценить пять линейных вкладов и часть парных взаимодействий, а остальные линейные вклады оценить вместе со взаимодействиями факторов. При этом остается достаточно степеней свободы для оценки точности статистических характеристик регрессии при высокой надежности выводов. Повторные опыты проводили в центре плана.

Реализацию ортогональных планов проводили для классической аттенюаторной регуляции экспрессии триптофана (ген *trpE*) и лейцина (ген *leuA*) у бактерии *Escherichia coli*. Для первой последовательности уже при исходных условиях работы модели получена значительная регуляция, а для второй при тех же условиях не было выявлено сколько-нибудь положительных результатов, в то время как в литературе имеются сведения, основанные на лабораторных данных, о наличии незначительной аттенюаторной регуляции.

Для первой последовательности реализация плана и регрессионный анализ результатов уже на первом этапе расширили факторную область, в которой имеется значительная регуляция, и позволили выделить пять из десяти факторов как значимо влияющие на увеличение регуляции. Для уточнения описания факторной области с хорошим качеством регуляции использовали новый план. Его условия и результаты реализации приведены в таблицах 1 и 2.

После удаления незначимых коэффициентов уравнение регрессии имеет вид:

$$\hat{y} = 44.31 - 2.06x_2 - 16.19x_3 + 2.19x_2x_3 + 3.31x_3x_5$$

В таблице 2 последний столбец содержит значения этой регрессии. Дисперсия адекватности для этого уравнения: 2.39, а дисперсия воспроизводимости по 10 повторным результатам в центре плана равна 2.08. По критерию Фишера гипотеза о его адекватности не отвергается при $\alpha = 0.05$.

Следует отметить, что изучаемая регуляция (ген *trpE*) в условиях этого плана имеет хорошее качество и устойчивый характер.

Анализ взаимодействий приведенного уравнения регрессии показал наличие двух областей факторного пространства, где качественную регуляцию получили при заметно разных значениях концентрации *c*. Таким образом, для триптофана в стационарной области всего три фактора с их взаимодействиями отвечают за качество регуляции. Причем остальные факторы фиксированы на уровнях, принятых в модели по умолчанию.

Таблица 1

Факторы	Условные и натуральные значения факторов		
x_1	-1	1	0
x_2	0.5	1.5	1
x_3	30	50	40
x_4	13.5	16.5	15
x_5	8	12	10

Таблица 2

№ опыта	x_2	x_3	x_4	x_5	y	$\hat{y}$
1	1	1	1	1	31	31.56
2	1	1	1	-1	25	24.94
3	1	1	-1	1	32	31.56
4	1	1	-1	-1	25	24.94
5	1	-1	1	1	52	52.94
6	1	-1	1	-1	60	59.56
7	1	-1	-1	1	52	52.94
8	1	-1	-1	-1	61	59.56
9	-1	1	1	1	33	31.3
10	-1	1	1	-1	27	24.68
11	-1	1	-1	1	29	31.3
12	-1	1	-1	-1	23	24.68
13	-1	-1	1	1	61	61.44
14	-1	-1	1	-1	66	68.06
15	-1	-1	-1	1	63	61.44
16	-1	-1	-1	-1	69	68.06

Исследование факторной области лейцина при той же стратегии потребовало большего количества шагов по реализации планов с последующим анализом результатов. В итоге была найдена факторная область, в которой устойчиво получали не столь значительную, как по триптофану, но весьма значимую регуляцию. Уравнение регрессии в этой области содержит факторы x_2, x_3, x_4.

III. Результаты симплексного планирования

Первая часть работы показала, что ортогональное планирование при поиске оптимальных условий требует большого объема эксперимента. Поэтому далее было использовано симплексное планирование [3] при варьировании одновременно пяти факторов, выделенных ранее на двух последовательностях (Таблица 1). Кроме того, решено было несколько изменить параметр *Y* для характеристики качества регуляции, приняв за него площадь под кривой вероятности на соответствующем участке. Воспроизводимость по такому параметру оказалась лучше, чем в первом варианте.

Поиск оптимальных условий при движении симплекса и запуск счета на модели хорошо алгоритмизуется и был программно реализован. Так было проанализировано более десятка различных последовательностей. В большинстве случаев обнаружена стационарная область со значительной регуляцией. В таблице 3 приведены для последовательности фенилаланина (ген *pheST*) координаты вершин начального симплекса в условных единицах и полученные значения *Y*, а в таблице 4 приведены условия и результаты движения симплекса.

Как видим, уже во второй вершине значение функции отклика много больше, почти 49 ед.

Таблица 3

	Координаты вершин начального симплекса						
№ опыта	x_1	x_2	x_3	x_4	x_5	x_6	Y
1	0	0	0	0	0	0	22.63
2	0.90	0.19	0.19	0.19	0.19	0.19	48.79
3	0.19	0.90	0.19	0.19	0.19	0.19	20.02
4	0.19	0.19	0.90	0.19	0.19	0.19	24.96
5	0.19	0.19	0.19	0.90	0.19	0.19	23.68
6	0.19	0.19	0.19	0.19	0.90	0.19	23.23
7	0.19	0.19	0.19	0.19	0.19	0.90	22.65

Таблица 4

	Вершины симплекса	x_1	x_2	x_3	x_4	x_5	x_6	Y
1		0	0	0	0	0	0	22.63
2		0.90	0.19	0.19	0.19	0.19	0.19	48.79
3		0.19	0.90	0.19	0.19	0.19	0.19	20.02
4		0.19	0.19	0.90	0.19	0.19	0.19	24.96
5		0.19	0.19	0.19	0.90	0.19	0.19	23.68
6		0.19	0.19	0.19	0.19	0.90	0.19	23.23
7		0.19	0.19	0.19	0.19	0.19	0.90	22.65
8	1 2 4 5 6 7 (3)	0.37	-0.58	0.37	0.37	0.37	0.37	23.46
9	2 4 5 6 7 8 (1)	0.68	0.13	0.68	0.68	0.68	0.42	44.17
10	2 4 5 6 8 9 (7)	0.65	-0.08	0.65	0.65	0.65	-0.38	42.91
11	2 4 5 8 9 10 (6)	0.80	-0.18	0.80	0.80	-0.14	0.14	51.31
12	2 4 5 9 10 11 (8)	0.77	0.73	0.77	0.77	0.22	-0.11	42.46
13	2 4 9 10 11 12 (5)	1.14	0.13	1.14	0.20	0.41	-0.04	43.02
14	2 9 10 11 12 13 (4)	1.45	0.11	0.51	0.90	0.48	-0.12	22.13
15	R_2 4 9 10 11 13 (12)	0.68	-0.60	0.68	0.13	0.44	0.29	48.36
16	2 9 10 11 13 15 (4)	1.13	0.01	0.58	0.40	-0.17	-0.28	57.40
17	2 9 11 13 15 16 (10)	1.13	-0.02	0.71	0.15	-0.18	0.62	48.73
18	2 9 11 15 16 17 (13)	0.63	-0.29	0.08	0.59	-0.13	0.50	33.93
19	R_2 11 13 15 16 17(9)	1.25	-0.28	0.69	-0.06	-0.50	-0.12	48.67
20	2 11 15 16 17 19(13)	0.82	-0.43	0.08	0.34	-0.53	0.32	40.57
21	2 11 16 17 19 20 (15)	1.33	0.36	0.34	0.48	-0.88	0.00	58.20
22	2 11 16 17 19 21 (20)	1.36	0.45	1.02	0.31	-0.03	-0.14	27.24
23	R_2 11 16 19 20 21(17)	0.95	-0.09	0.18	0.57	-0.50	-0.53	53.01
24	2 11 16 19 21 23 (20)	1.30	0.43	0.85	0.45	-0.14	-0.52	37.25
25	R_2 11 16 17 20 21 (19)	0.79	0.26	0.21	0.85	-0.07	0.45	45.19

Далее осуществлялась пошаговая реализация методики, при которой вершина с худшим результатом заменялась противоположной (согласно алгоритму и программе). В таблице 4 показаны условия, движение симплекса и результаты счета на модели в каждой вершине симплекса. В первом столбце указан номер опыта. Во втором столбце указаны номера вершин, участвующих в опыте. В скобках указана вершина с наименьшим значением функции отклика, которую исключили, рассчитывая координаты новой вершины. Символ «R_» указывает на то, что полученное значение функции отклика оказалось меньше, чем полученное значение той же функции в предыдущем опыте, в связи с чем оно было отброшено. В последующих столбцах указаны значения факторов в условных единицах и значение функции отклика.

Счет на модели при замене вершин симплекса привел к повороту симплекса относительно вершины с результатами по функции отклика 54 – 58 ед. Последующие опыты не привели к дальнейшему улучшению результата.

Корреляционный анализ полученных результатов в стационарной области показал, что таблица парных коэффициентов корреляции пригодна для выделения связей между факторами и их взаимодействиями. Это позволяет выделять значимые связи факторов и их взаимодействий и адекватно описать стационарную область уравнением регрессии. Такой анализ был проведен для десяти последовательностей гена *pheST* таксономической группы α-протеобактерий, а также гена *pheS* таксономической группы γ-протеобактерий.

IV. Выводы

Исследование показало обоснованность и успешность применения регрессионного анализа на основе ортогонального планирования эксперимента для изучения математической модели классической аттенюаторной регуляции экспрессии генов у бактерий и, возможно, других математических моделей в области регуляции и эволюции.

При поиске оптимальных условий наиболее эффективным следует признать симплексное планирование, которое можно проводить автоматически, используя готовую программу.

В случае, когда результат счета модели не является числом, возможно ввести параметр, характеризующий результат счета с помощь некоторой обобщенной числовой оценки для эмпирической оптимизации исследования.

Выделены для разных последовательностей до пяти основных параметров модели (x_1, …, x_5), позволяющих с высокой точностью описать их связь с параметром Y, характеризующим качество регуляции. При этом другие параметры фиксированы на указанных в модели уровнях, а их влияние в стационарной факторной области статистически незначимо.

На основе регрессионного анализа получена последовательность значений параметров модели для большого числа объектов, быстро сходящаяся к эмпирическим значениям, которые были выбраны по умолчанию в модели [1] и ее программной реализации RNAnodel.

Библиография

[1] V. A. Lyubetsky, S. A. Pirogov, L. I. Rubanov, A. V. Seliverstov, “Modeling classic attenuation regulation of gene expression in bacteria,” *Journal of Bioinformatics and Computational Biology*, vol. 5, no 1, pp. 155–180, 2007.

[2] С. Н. Истомина, Л. И. Рубанов, “Параллельный алгоритм поиска регуляторного сигнала в геномах бактерий”, *Информационные процессы*, т. 2, ч. 1, 2002.

[3] К. Хартман, Э. Лецкий, В. Шефер, *Планирование эксперимента в исследовании технологических процессов*, М.: Мир, 1977.

Optimization of the factor region of the gene expression regulation model

Svetlana. N. Istomina

Abstract — **In this study we test parameters of the gene expression regulation model to identify key factors affecting predictions. The task is to find conditions to better describe regulations, for which we made orthogonal and simplex planning of the experiment. Higher reliability of predictions was ensured by using multidimensional correlation analysis and regression equations that adequately describe the stationary region and take into account interactions of second-order factors.**

Keywords — **mathematical model, expression, regulation, genes, optimal planning, orthogonality, simplex, correlation and regression analyses.**