БИОФИЗИКА, 2017, том 62, вып. 6, с. 1054–1066

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 57.087

МЕТОД АНАЛИЗА ПРЕДСКАЗАТЕЛЬНОЙ СИЛЫ МОДЕЛИ БИОЛОГИЧЕСКОЙ СИСТЕМЫ С НИЗКОЙ ЧУВСТВИТЕЛЬНОСТЬЮ К ПАРАМЕТРАМ

© 2017 г. Е.М. Мясникова, А.В. Спиров*

Санкт-Петербургский политехнический университет Петра Великого, 195251, Санкт-Петербург, Политехническая ул., 29

**Институт эволюционной физиологии и биохимии им. И.М. Сеченова РАН, 194223, Санкт-Петербург, просп. Тореза, 44*

E-mail: myasnikova@spbcas.ru

Поступила в редакцию 20.02.17 г.

Неоднозначность оценок параметров модели биологической системы может являться следствием низкой чувствительности модели к пертурбациям входных данных (параметров), что математически отражает биологические механизмы робастности. Нами разработан новый метод оценки предсказательной силы модели в условиях неопределенности оценивания параметров. Под предсказаниями мы понимаем правильное воспроизведение моделью поведения системы при изменении входных данных и параметров. Метод основан на анализе относительной чувствительности решения, полученного подгонкой к данным, к параметрам предсказываемой модели. Наш подход продемонстрирован на примере модели формирования паттерна экспрессии мРНК гена *hb* у эмбриона дрозофилы и ее способности предсказания паттерна hb в нуль-мутанте по гену *Kr*. Нелинейный характер системы моделируется насыщающей сигмоидной функцией, что является причиной низкой чувствительности. Метод позволяет оценить предсказательную силу модели и вскрыть причины плохих предсказаний, а также выбрать релевантный с точки зрения предсказаний уровень детализации модели.

Ключевые слова: математическая модель, биологическая система, анализ чувствительности, предсказательная сила.

Типичной проблемой, возникающей при моделировании биологических систем, является неоднозначность в определении оценок параметров модели. В задачах системной биологии среди параметров часто встречаются такие, к которым система очень мало чувствительна, т.е. при изменении их значения на несколько порядков качество подгонки практически не меняется (в англоязычной литературе такие параметры называются “sloppy”). Подобная нечувствительность представляет собой важное свойство модели, которое математически выражает способность системы демонстрировать робастность, устойчивость к изменениям окружающей среды и внутренним возмущениям без потери функциональности.

В то же время другой источник неоднозначности (неидентифицируемости) оценок – сильные корреляции между оценками параметров – может являться следствием неправильной параметризации или сильной зашумленности данных. Все эти типы неопределенности могут приводить к низкой предсказательной силе моделей. Для количественной характеристики предсказательных свойств модели используем методы анализа чувствительности и идентифицируемости.

Анализ чувствительности – методология, позволяющая оценить, насколько изменение значений параметров (входных данных) влияет на изменение результатов моделирования. Это позволяет выявить параметры и входные данные, которые вносят наибольший и наименьший вклад в динамику системы [1–4]. Такие параметры будем соответственно называть «жесткими» (stiff) и «слабыми» (sloppy). Если перед нами не стоит задача нахождения оценок отдельных параметров, то поведение системы может быть исчерпывающе описано при помощи только жестких параметров (или комбинаций параметров), т.е. тех, к которым система наиболее чувствительна, при этом не принимая в расчет не интересующие нас слабые параметры [1,5].

Сокращения: RSS – relative stiff sensitivity, RSM – relative sensitivity measure, RCM – relative correlation measure.

Обычно локальный анализ практической неидентифицируемости, являющейся следствием недостаточности данных, проводится на основе исследования матрицы чувствительности в окрестности решения [6,7]. Важным свойством сложных систем является то, что для хороших предсказаний достаточно обеспечить идентифицируемость комбинаций жестких параметров модели [1,5].

В целом правильное предсказание динамики системы требует надежных и единственных оценок жестких параметров, определяющих поведение предсказываемой модели. Если же при подгонке модели выявляется низкая чувствительность к этим параметрам, предсказание может оказаться неверным.

Помимо того, если имеются сильные корреляции между жесткими параметрами предсказываемой модели и остальными параметрами, определяемыми подгонкой, предсказание может оказаться некорректным. Это происходит из-за того, что в этом случае изменение значений одних параметров влечет за собой изменения и в других, коррелированных с ними. Таким образом, изменения параметров взаимно компенсируют эффект, оказываемый на значения целевого функционала, и, следовательно, если значение части из них изменены, то оценки остальных могут оказаться ненадежными.

Цель нашей работы состоит в том, чтобы охарактеризовать модель и ее конкретные решения с точки зрения предсказательных свойств. Это позволит выбрать среди вариантов моделей и их допустимых решений те, которые обеспечивают наилучшие предсказания. Особое внимание уделяется нечувствительным моделям и их предсказательным свойствам.

С этой целью нами разработан метод оценки предсказательной силы модели в условиях неопределенности оценивания параметров. Идея, лежащая в основе метода – сравнение чувствительности двух моделей: модели, решение которой определено подгонкой к данным эксперимента (*полная модель*), и модели, поведение которой при изменении входных данных мы намерены предсказать (*предсказываемая модель*). Модели главным образом характеризуются каждая своими жесткими комбинациями параметров, и чувствительность подгоняемой модели к жестким комбинациям параметров предсказываемой модели может служить характеристикой способности модели к точным предсказаниям.

Мы тем самым выявляем источники плохих предсказаний, связанные с недостаточностью информации, причем наш метод позволяет классифицировать источники и степень неопределенности предсказаний. Также наш подход позволяет определить, как усложнение модели за счет включения дополнительных входных данных и параметров влияет на ее предсказательные свойства.

Мы изучаем проблемы предсказания на примере модели регуляции системы генов сегментации эмбриона дрозофилы. Система генов представляет собой генную сеть, т.е. совокупность координировано экспрессирующихся генов, их белковых продуктов и взаимосвязей между ними. Регуляторные взаимодействия осуществляются посредством связывания продуктов генов системы (белков), называемых транскрипционными факторами, с определенными участками последовательности ДНК гена (сайтами связывания). Таким образом, состояние системы характеризуется экспрессией гена (способностью производить РНК и белок) в каждой клетке/ядре и динамически определяется концентрациями регулирующих их транскрипционных факторов. Динамика системы многократно успешно моделировалась на разном уровне детализации [8–13]. Регуляция генов в этих моделях описывается нелинейной насыщающей функцией сигмоидного типа. Вследствие свойства насыщения модель может обладать низкой чувствительностью к параметрам. Предсказательная сила модели здесь будет пониматься как способность правильно воспроизводить поведение предсказываемой системы при изменении входных данных. Часто проблема может быть сформулирована в терминах чувствительности модели к определенным параметрам модели и их идентифицируемости. Например, нуль-мутанты (т.е. эмбрионы, в которых отсутствует экспрессия одного/нескольких генов) моделируются за счет обнуления параметров, относящихся к отсутствующему гену.

Решения модели, обладающие и не обладающие свойствами насыщения, имеют различные предсказательные свойства, что демонстрируется с использованием нашего метода.

ОТНОСИТЕЛЬНЫЕ МЕРЫ ЧУВСТВИТЕЛЬНОСТИ

Корректное предсказание поведения системы является необходимым качеством математической модели. Нас интересуют предсказания поведения решений модели при изменении некоторых входных данных, которые можно охарактеризовать изменением значений соответствующих параметров. Наш подход основан на локальном анализе чувствительности модели к параметрам.

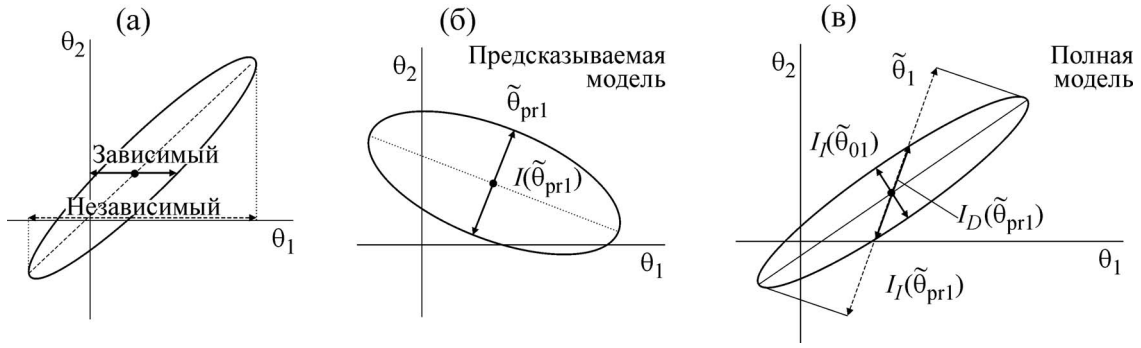

Рис. 1. Построение относительных мер чувствительности модели к параметрам в двумерном случае. (а) – Зависимый и независимый доверительные интервалы для параметра θ_1 показаны в виде двусторонних стрелок. (б) – Доверительная область предсказываемой модели: $\tilde{\theta}_{pr,1}$ – первая главная компонента, жесткий параметр; серая стрелка – доверительный интервал для $\tilde{\theta}_{pr,1}$. (в) – Доверительная область полной модели: $\tilde{\theta}_{01}$ – первая главная компонента; черная сплошная стрелка – зависимый, а пунктирная стрелка – независимый доверительный интервал для $\tilde{\theta}_{pr,1}$ в полной модели. Пусть $L = 1$, тогда $RSS = I^2(\tilde{\theta}_{pr,1})/I_D^2(\tilde{\theta}_{pr,1})$; $RSM = I^2(\tilde{\theta}_1)/I_D^2(\tilde{\theta}_{pr,1})$ и $RCM = I^2(\tilde{\theta}_{pr,1})/I_I^2(\tilde{\theta}_{pr,1})$.

Источником низкой предсказательной силы может быть: 1) низкая чувствительность функционала к параметрам, которые определяют предсказываемое поведение системы; 2) наличие сильных корреляций между фиксированными параметрами и теми, значения которых определяются в результате подгонки. В качестве критерия предсказательной силы использованы показатели, учитывающие вышеуказанные характеристики параметров модели.

Нами предложены два вида количественной характеристики относительной чувствительности *RSS* (Relative Stiff Sensitivity) и *RSM* (Relative Sensitivity Measure).

В предположении адекватности модели обе меры характеризуют локальную возможность получения оценки подмножества параметров, достаточно точной для предсказаний. Подход основан на анализе свойств жестких комбинаций параметров, которые определяются для обеих моделей: подгоняемой и предсказываемой.

Пусть биологическая система описывается системой обыкновенных дифференциальных уравнений, параметры модели оцениваются путем подгонки к экспериментальным данным минимизацией целевого функционала среднеквадратичной разности между решениями и наблюдениями. Ошибки измерения предполагаются нормально распределенными.

Чувствительность модели характеризуется формой и размерами доверительной области всего набора параметров θ, которая строится с помощью матрицы чувствительности $\mathbf{J}$ в окрестности решения $\hat{\theta}$:

$$(\theta - \hat{\theta})^T \mathbf{J}^T \mathbf{J} (\theta - \hat{\theta}) \le R^2, \qquad (1)$$

где R – константа, зависящая от числа параметров и точек наблюдения, оптимального значения функционала и уровня доверия.

Элементами матрицы $\mathbf{J}$ являются частные производные решения модели по параметрам, вычисленные в точках наблюдений. Примеры доверительной области показаны на рис. 1.

Характеристикой чувствительности модели к параметрам и их линейным комбинациям могут служить два вида доверительных интервалов: зависимые $I_D(\theta)$ и независимые $I_I(\theta)$ [7]:

$$I_D(\theta_i) = R/\sqrt{(\mathbf{J}^T\mathbf{J})_{ii}};\ I_I(\theta_i) = R\sqrt{(\mathbf{J}^T\mathbf{J})^{-1}_{ii}}. \qquad (2)$$

Геометрически они определяются как, соответственно, сечение и проекция доверительной области, параллельные оси параметра (комбинации параметров), как показано на рис. 1а. Оба вида доверительных интервалов характеризуют точность оценки параметра – чем короче доверительные интервалы, тем точнее оценка. При этом зависимый доверительный интервал характеризует чувствительность модели к параметру независимо от остальных параметров, а независимый учитывает также корреляции между параметрами. При этом оба доверительных интервала совпадают только в случае некоррелированности параметров. Все сказанное справедливо также и для комбинаций параметров.

Жесткие комбинации параметров выделяются методом главных компонент и соответствуют на графике главным осям эллипсоида (см. работу [14]). Для этих комбинаций оба вида доверительных интервалов совпадают ввиду их независимости $I_I(\tilde{\theta}) = I_D(\tilde{\theta}) = I(\tilde{\theta})$ (рис. 1а).

Пусть в результате подгонки получено решение полной модели, характеризующееся вектором параметров θ_0, и нас интересует поведение модели при измененных значениях некоторых из этих параметров, т.е. при векторе параметров θ_{pr}. Пусть из θ_0 и θ_{pr} выделены жесткие комбинации параметров (главных компонент) для полной и предсказываемой модели, которые обозначаем $\tilde{\theta}_0$ и $\tilde{\theta}_{pr}$ соответственно. Несколько таких комбинаций (которым соответствуют самые короткие доверительные интервалы) главным образом определяют поведение каждой из моделей. Для краткости будем их называть просто *жесткими параметрами*. Следует отметить, что параметры $\tilde{\theta}_{pr}$ не являются главными компонентами для полной модели и их зависимые и независимые доверительные интервалы могут не совпадать.

В качестве меры чувствительности модели к вектору параметров θ будем рассматривать сумму обратных значений квадратов длин зависимых доверительных интервалов

$$Sens(\theta) = \sum_{i=1}^{L} I_D^{-2},$$

где L – число информативных главных компонент. Определим *жесткую чувствительность* как чувствительность модели к своим жестким параметрам

$$SSens(\theta) = Sens(\tilde{\theta}) = \sum_{i=1}^{L} I^{-2}(\tilde{\theta}_i). \quad (3)$$

Мера RSS определяется как отношение чувствительностей обеих моделей к жестким параметрам предсказываемой модели $\tilde{\theta}_{pr}$:

$$RSS = \frac{SSens(\theta_{pr})}{Sens(\tilde{\theta}_{pr})} = \frac{\sum_{i=1}^{L} I^{-2}(\tilde{\theta}_{pr,i})}{\sum_{i=1}^{L} I_D^{-2}(\tilde{\theta}_{pr,i})}. \quad (4)$$

Этот показатель отражает количество информации о параметрах предсказываемой модели, содержащееся в подгоняемой модели, по сравнению с жесткой чувствительностью самой предсказываемой модели. Показатель RSS принимает значения близкие к нулю, если предсказываемая модель высокочувствительна к своим жестким параметрам, а при подгонке полной модели именно эти параметры несущественны, т.е. их оценки неточны. Высокие значения RSS (близкие к единице и выше) свидетельствуют о том, что диапазон значений параметров, необходимый для предсказаний, широк и полученные в результате подгонки оценки будут достаточно точными.

Однако мера RSS обладает тем недостатком, что ее значение не определено в случае нечувствительности к своим параметрам обеих моделей: полной и предсказываемой, что приводит к очень малым значениям, как числителя, так и знаменателя в выражении (4). В такой ситуации более информативным является показатель

$$RSM = \frac{Sens(\tilde{\theta}_{pr})}{SSens(\theta_0)} = \frac{\sum_{i=1}^{L} I_D^{-2}(\tilde{\theta}_{pr,i})}{\sum_{i=1}^{L} I^{-2}(\tilde{\theta}_{0,i})}; \quad (5)$$

т.е. отношение жесткой чувствительности (3) предсказываемой модели к жесткой чувствительности полной модели. RSM может принимать значения от нуля до единицы, причем близкие к нулю значения означают, что на фоне своих параметров полная модель низкочувствительна к жестким параметрам предсказываемой модели, оценки которых поэтому могут оказаться недостаточно точными для правильного предсказания поведения этой модели.

В то же время если полная модель слабо чувствительна ко всем параметрам (в том числе и к их жестким комбинациям), то она в любом случае будет слабо чувствительна и к параметрам предсказываемой модели, т.е. значения RSM не будут отражать точности предсказаний. В описываемой ситуации более информативным показателем является RSS.

До сих пор мы не учитывали корреляции между параметрами. Этот источник неопределенности модели характеризуется соотношением зависимых и независимых доверительных интервалов и определяется как

$$RCM = \frac{\sum_{i=1}^{L} I_I^{-2}(\tilde{\theta}_{pr,i})}{\sum_{i=1}^{L} I_D^{-2}(\tilde{\theta}_{pr,i})}. \quad (6)$$

Показатель RCM (Relative Correlation Measure) принимает значения в диапазоне от нуля до единицы, причем малые значения свидетельствуют о наличии сильных корреляций и, следовательно, неидентифицируемости жестких параметров $\tilde{\theta}_{pr}$ при подгонке полной модели, что также может приводить к плохим предсказаниям.

В итоге близкое к нулю значение одной из двух мер относительной чувствительности (RSS

или RSM), а также меры RCM свидетельствует о невозможности точных предсказаний. На основе значений этих показателей можно сделать вывод о предсказательных свойствах модели и определить источники плохих предсказаний. Пример построения всех мер показан на рис. 1 для случая двух параметров.

Ограничения метода. Как любой способ анализа чувствительности, метод требует очень тщательной нормировки параметров, т.е. все параметры должны быть сведены к одному порядку значений.

Ввиду локальности метода свойства решения при изменении вектора параметров при предсказаниях могут довольно существенно меняться. Это, например, проявляется в вышеназванной ситуации, когда наряду с полной моделью предсказываемая модель также оказывается нечувствительной к параметрам, т.е. мы попадаем в окрестность другого «слабого» решения. Именно поэтому мы вводим две разные меры, каждая из которых выявляет низкую относительную чувствительность в разных ситуациях.

Надо иметь в виду, что предсказанное решение не обязательно является оптимальным, которое можно было бы получить методом подгонки к данным предсказываемого эксперимента, если бы они нам были доступны. Ввиду этого, мы можем только выбирать из решений, обеспечивающих хорошее согласие с имеющимися данными, те, которые по своим свойствам чувствительности могут быть использованы для предсказаний.

Строго говоря, мера RSS вычисляется с точностью до множителя, так как она определяется в предположении, что коэффициент R в выражении (1) одинаков для обеих моделей, что верно только при одинаковой вариабельности данных, используемых для подгонки и предсказаний. Однако очень малые значения RSS оказываются информативными для широкого спектра экспериментальных ситуаций.

РЕЗУЛЬТАТЫ

Модель с насыщением. Далее покажем применение нашего метода на примере нескольких вариантов модели одной и той же простой биологической системы. Выбираем те решения, которые хорошо описывают систему и в то же время наглядно демонстрируют типичные проблемы с предсказаниями вследствие низкой чувствительности к параметрам. Рассмотрим следующую ситуацию. Пусть нелинейная динамика модели описывается сигмоидной функцией, обладающей свойством насыщения, т.е. при значениях аргумента, превышающих определенный порог, значения функции практически не изменяются (рис. 2а). Это означает нечувствительность модели к параметрам в окрестности такого решения, а следовательно, их локальную неидентифицируемость.

Теперь представим себе, что при подгонке получен вектор оценок параметров, при котором достигается насыщение. Нас интересует, сможем ли мы правильно предсказать поведение решения модели при изменении значений некоторых из параметров. Однако может случиться так, что после такого изменения аргумент сигмоиды окажется вне зоны насыщения, т.е. решение станет вполне чувствительно в окрестности новых значений параметров и для корректного предсказания потребуется их идентифицируемость. Следовательно, корректное предсказание окажется невозможным.

Описание модели. Проиллюстрируем принцип применения предложенного нами критерия на примере простой модели, полное описание которой приводится в работе [13]. Модель описывает формирование паттернов экспрессии гена hunchback (*hb*) у эмбриона дрозофилы во времени и в пространстве с учетом регуляторной структуры ДНК, контролирующей экспрессию. Паттерн экспрессии представляет собой распределение мРНК данного гена по ядрам эмбриона (рис. 3), в нашем случае в одномерной цепочке из 100 ядер вдоль его продольной оси. Два транскрипта мРНК hb продуцируются с двух промоторов P1 и P2 (регуляторных элементов ДНК, отвечающих за старт транскрипции). Оба транскрипта транслируются в один и тот же белок Hb, но формируют различные независимые паттерны, что позволяет нам рассматривать экспрессию с двух промоторов отдельно. Суммарная экспрессия Hb с двух промоторов формирует общий выход модели.

Паттерны экспрессии обоих промоторов моделируются во времени и пространстве (ядрах эмбриона) уравнением реакции диффузии $\partial x_t/\partial t = Rg(x_t) - \lambda x_t + D\cdot\mathrm{diff}(x_t)$, в котором синтез РНК описывается членом реакции:

$$Rg(\Phi(TF_1, TF_2, ..., TF_k) + H), \qquad (7)$$

где $g(x) = x/\sqrt{x^2+1}$ при $x \geq 0$, $g(x) = 0$ при $x < 0$ – сигмоидная функция, R – коэффициент синтеза, H – порог синтеза. Вид функционала Ф задается в рамках модели и представляет собой сумму в общем случае нелинейных членов от концентрации транскрипционных факторов – продуктов генов-регуляторов TF_k. Резкие границы областей экспрессии динамически

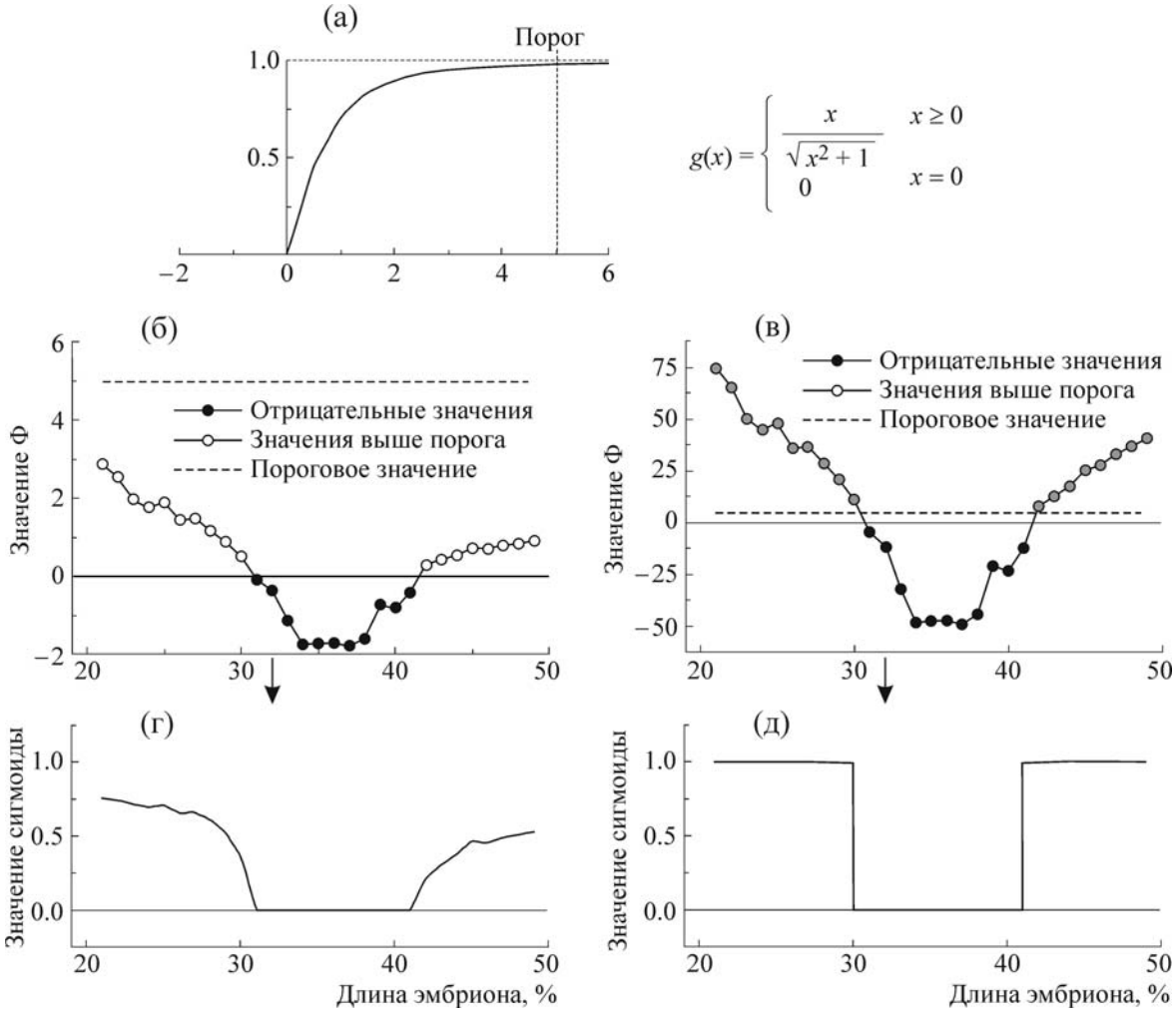

Рис. 2. Механизм насыщения. (а) – Функция сигмоиды. (б)–(д) – Примеры решений без насыщения – (б) и (г) – и с насыщением (ступенчатое) – (в) и (д). Значения аргумента функции $g(\cdot)$ обозначены кружками на рис. (б) и (в), они отображаются в значения сигмоиды (сплошная линия на рис. (г) и (д)). Значения, превышающие пороговое (= 5, пунктирная линия), показанные серыми кружками, отображаются в значение сигмоиды, почти равное единице; отрицательные значения, показанные черными кружками, отображаются в ноль. Остальные значения аргумента обозначены белыми кружками, в этих точках нет насыщения и $0 < g(x) < 1$.

сглаживаются за счет членов диффузии и распада уравнения реакции-диффузии.

Мы изучаем поведение системы у двух генотипов дрозофилы, а именно эмбрионов дикого типа, т.е. немутантных организмов, а также нуль-мутантов по гену Kruppel (*Kr*), т.е. тех, в которых отсутствует экспрессия гена *Kr* (*Kr*$^-$-мутантах). Мутант моделируется обнулением входных данных по экспрессии *Kr*, а также изменением входных данных по другим генам (известных из эксперимента [15]). В модели акцент делается на экспрессии пика «PS4» в центральной части эмбриона, которая в эмбрионах дикого типа делится в некоторой пропорции между паттернами P1 и P2, и при этом значительно снижается в эмбрионах, нуль-мутантных по *Kr* (рис. 3а,б) [16,17]. Мы ставим своей целью оценить способность модели воспроизводить паттерн экспрессии hb на уровне мРНК в *Kr*$^-$-мутантах.

Пример 1. Решения модели с насыщением и без насыщения и их предсказательная сила. Известно, что в мутантном эмбрионе экспрессия с промотора P1 отсутствует [16,17] и, следовательно, модель P1 должна при обнулении входных данных по экспрессии *Kr* или, что то же самое, при обнулении параметров, описывающих воздействие *Kr* (*Kr*-параметров), приводить к исчезновению пика. В работе [13] предложены три версии модели, в которых этот эффект достигается как за счет математического вида уравнения, так и на биологической основе. Во всех случаях при обнулении *Kr*-параметров вы-

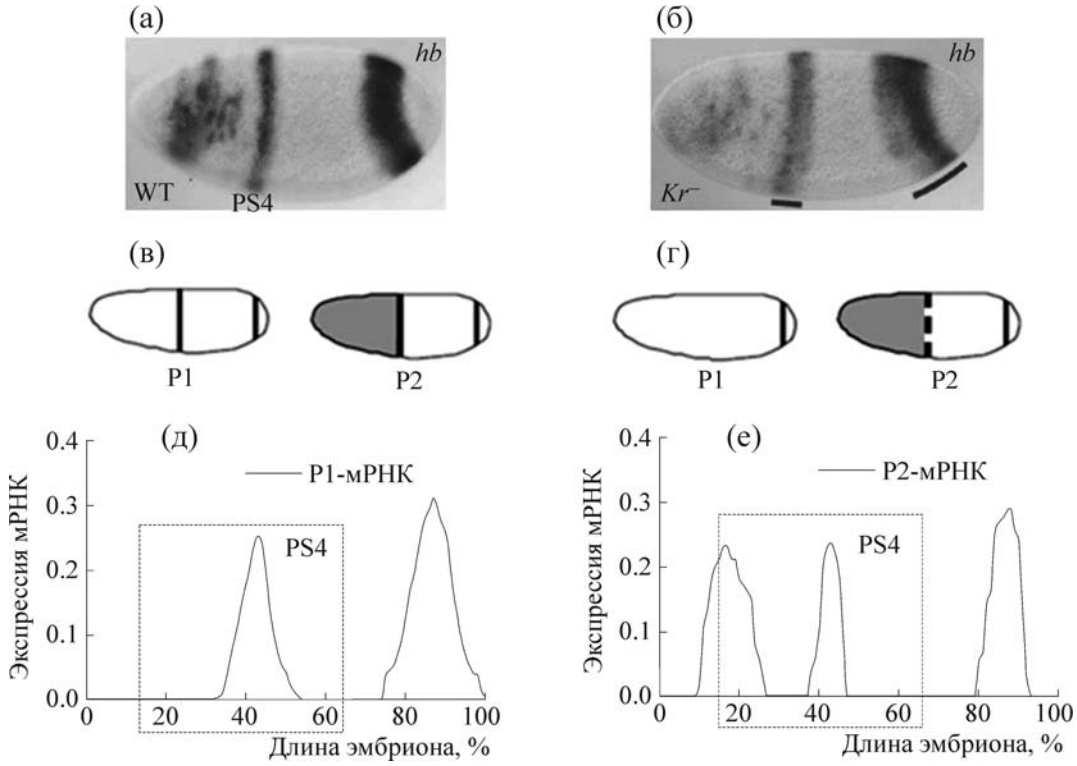

Рис. 3. Паттерн экспрессии мРНК *hb* в эмбрионах дикого типа и Kr⁻. (а), (б) – Конфокальные изображения суммарной экспрессии *hb* [16]. Видно, что у мутанта полоса PS4 редуцирована. (в), (г) – Схема распределения транскрипта в диком типе (а) и мутанте (б) соответственно. Вертикальная сплошная линия соответствует полноразмерному пику, пунктирная – редуцированному. (д), (е) – Примеры графиков количественных данных, использованных для подгонки. Дикий тип: паттерн экспрессии P1 (в) и P2 (г). Пунктиром выделена область моделирования.

ход модели P1 обнуляется. Мы рассмотрим один из вариантов функционала Ф для P1:

$$\Phi_1: \theta^1_{Kr1}[\mathrm{Kr}] - \theta^1_{Kr2}[\mathrm{Kr}]^2 - \\ - \theta^1_{Gt}[\mathrm{Gt}] - \theta^1_{Kni}[\mathrm{Kni}] - \theta^1_{Nub}[\mathrm{Nub}], \quad (8)$$

т.е. экспрессия P1 регулируется четырьмя транскрипционными факторами – Kr, Gt, Kni и Nub (в квадратных скобках – концентрация транскрипционного фактора в ядре).

Паттерн экспрессии промотора P2 задается единым уравнением

$$\Phi_2: \theta_{Bcd}F([\mathrm{Bcd}]; K_{Bcd}) - \theta_{Kr}F([\mathrm{Kr}]; K_{Kr}) - \\ - \theta_{Kni}[\mathrm{Kni}] - \theta_{Gt}[\mathrm{Gt}] - \theta_{Nub}[\mathrm{Nub}], \quad (9)$$

где $F([\mathrm{TF}];K_{TF}) = [\mathrm{TF}]/(K_{TF} - [\mathrm{TF}])$ и определяется регуляторным воздействием трех транскрипционных факторов (Bcd, Kr и Gt). Уравнения (8) и (9) определяют экспрессию в каждом ядре. При обнулении *Kr*-параметров паттерн P2 изменяется в области экспрессии Kr, что выражается в уменьшении высоты пика PS4. Выделим из всего множества параметров отдельно параметры, отвечающие за формирование паттернов P1 и P2. Обозначим эти подмножества θ_1 и θ_2 соответственно. Те же подмножества с исключенными *Kr*-параметрами обозначим θ_1^{Kr} и θ_2^{Kr}. Таким образом, полная модель в данном случае определяется параметрами $\theta_1 \cup \theta_2$, оценки которых получены подгонкой, а предсказываемая модель мутанта – параметрами $\theta_1^{Kr} \cup \theta_2^{Kr}$.

Подгонка моделей производится к суммарной мРНК (P1 + P2). Точных количественных данных по экспрессии отдельно с каждого промотора не имеется, единственное свидетельство о том, что P2 экспрессируется в области PS4, имеется в работе [17]. Однако определить относительный вклад в экспрессию PS4 каждого из промоторов по приведенным в этой статье изображениям эмбриона не представляется возможным. Также неизвестно точно, меняется ли ширина домена PS4 в *Kr*⁻-мутанте по сравнению с диким типом.

Исходя из этого и представления о том, что пик PS4 в *Kr*⁻-мутантах сильно редуцирован [15–17], проанализируем предсказательные свойства модели. Рассмотрим разные решения

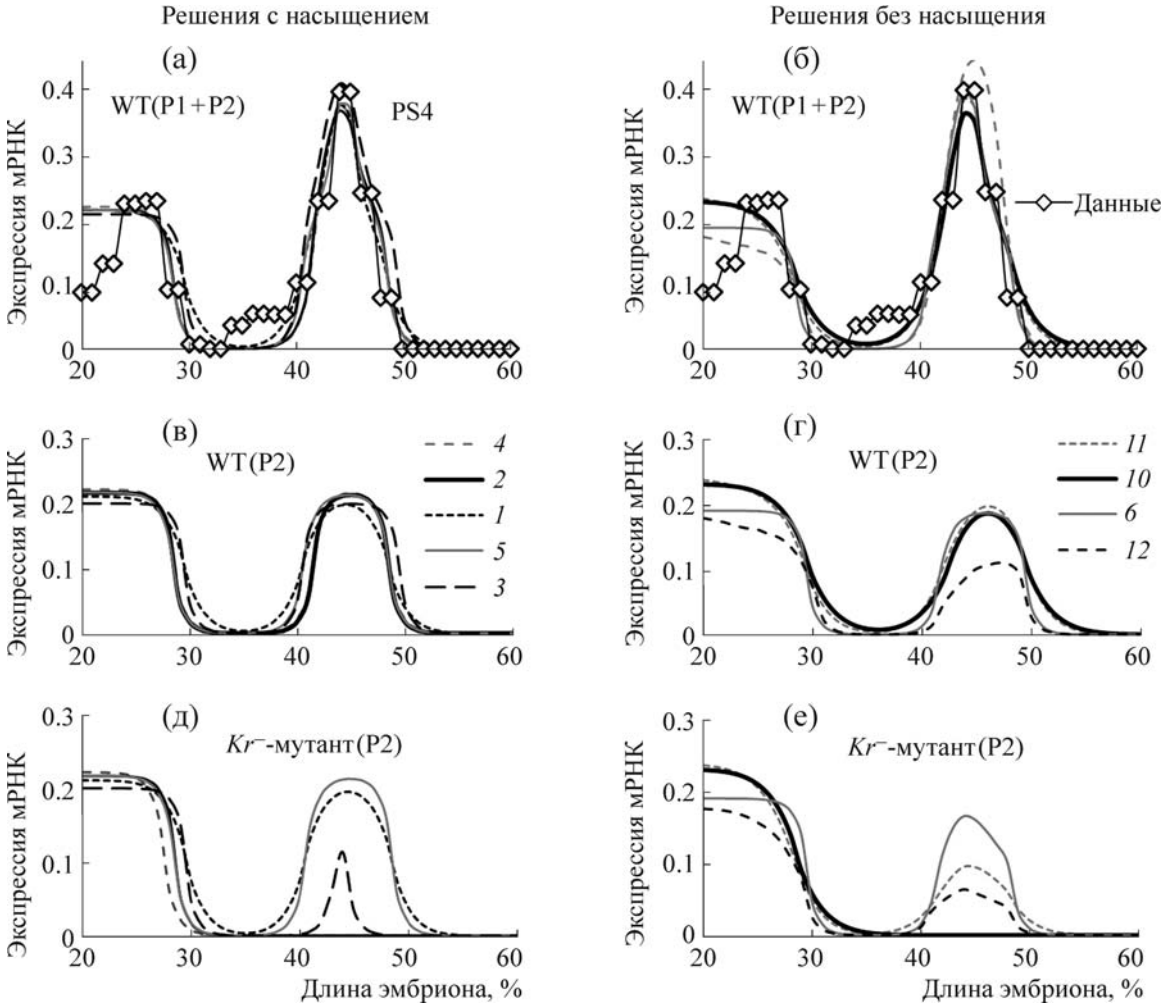

Рис. 4. Пример 1. Решения модели, полученные в результате подгонки и предсказаний. (а), (в), (д) – с насыщением; (б), (г), (е) – без насыщения. Данные по экспрессии мРНК показаны белыми ромбами; решения модели пронумерованы согласно табл. 1. (а), (б) – модель с подгонкой к суммарной мРНК (P1 + P2) в диком типе (WT); (в), (г) – модель с подгонкой к экспрессии P2 в диком типе; (д), (е) – предсказанная модель P2 в мутантах по *Kr*.

для дикого типа, воспроизводящие вклады промоторов в формирование PS4 в разных пропорциях, при этом обеспечивающие хорошее соответствие суммарной мРНК.

При разных соотношениях P1 и P2 наблюдается разное поведение модели. В силу сигмоидного вида уравнений модели максимальное значение пика PS4 в модели P2 ограничивается высотой антериорного (первого слева) пика экспрессии *hb* (рис. 4в). Такое решение достигается при насыщении сигмоидной функции в уравнении P2 (т.е. реализуется схема из рис. 2в,д), что выражается в низкой чувствительности к параметрам θ_2. Обнуление *Kr*-параметров в таком решении может либо не изменить вида паттерна, что нереалистично с биологической точки зрения, либо уменьшить его амплитуду (вплоть до нуля), так что аргумент сигмоиды не достигнет насыщения (схема на рис. 2б,г). В этом случае мы оказываемся в ситуации, описанной выше: относительная чувствительность полной модели к жестким комбинациям параметров модели мутанта θ_2^{Kr} оказывается очень низкой (обе меры *RSS* и *RSM* принимают значение, близкое к нулю), параметры θ_2^{Kr} неидентифицируемы, слабые и точное предсказание неосуществимо. Это выражается, например, в том, что идентичные решения для дикого типа приводят к совершенно различным предсказаниям в мутантах (рис. 4в,д). Другая возможная ситуация такова: оба решения P2 как в диком типе, так и в мутанте порождают пик PS4 высотой ниже максимальной (рис. 4г,е), т.е. решения не достигают насыщения (схема на рис. 2б,г). Тогда по подгонке дикого типа получаем более высокую чувствительность ко

Таблица 1. Пример 1. Значения мер чувствительности

	RSM	*RSS*	*RCM*	*RSM*	*RSS*	*RCM*	*RSM*	*RSS*	*RCM*
1	**1,8E–05**	2,0E–03	0,01	0,22	**0,007**	0,002	0,04	0,54	0,99
2	**3,2E–05**	N/A	0,05	0,53	**N/A**	1,0E–04	3,4E–05	1,00	0,03
3	**4,5E–05**	1,0E–06	0,00	0,34	**3,3E–05**	0,0007	0,13	0,87	0,43
4	**4,9E–05**	0,00	0,05	0,42	**0,99**	1,0E–05	0,02	0,54	0,98
5	**0,03**	0,01	0,01	0,91	**0,07**	0,005	0,76	0,58	0,92
6	0,12	0,00	1,0E–04	0,98	0,98	0,02	1,00	0,63	0,86
7	0,65	0,98	3,0E–05	0,92	0,70	2,0E–03	0,67	2,90	0,39
8	0,99	0,12	0,008	0,95	0,98	1,0E–03	1,00	0,67	0,8
9	0,99	0,57	0,001	1,00	0,71	9,0E–04	1,00	1,16	0,52
10	0,91	N/A	0,00	0,71	N/A	1,0E–07	N/A	N/A	N/A
11	0,97	0,20	0,0004	0,3	0,69	3,0E–04	1,00	0,75	0,73
12	0,99	1,06	0,01	1,00	0,65	0,006	1,00	1,70	0,36

Примечание. Жирным выделены примеры насыщения. Упорядочено по порядку убывания высоты пика PS4 в паттерне P2 дикого типа. Меры вычисляются в области PS4 от 30 до 65% EL. Первые три столбца – подгонка по множеству параметров $\theta_1 \cup \theta_2$ к суммарной РНК (P1 + P2); средние три столбца – подгонка по множеству параметров θ_2 к паттерну экспрессии P2; последние три столбца – подгонка по θ_2 к паттерну P2 в диком типе и мутанте.

всем параметрам (в том числе к θ_2^{Kr}) и более однозначное предсказание. Эта тенденция отражается в значениях меры относительной чувствительности *RSM*. В целом можно наблюдать явную корреляцию между высотой пика PS4 в диком типе (относительно первого пика P2) и значением *RSM*. Чем ниже PS4, тем меньше информации о параметрах теряется за счет насыщения, тем надежнее оценки и тем выше *RSM*.

Понятно, что в силу неоднозначности соотношения решений P1 и P2 оценки параметров θ_1 и θ_2 всегда будут коррелированы и часть параметров неидентифицируема, что отражается в невысоких значениях *RCM*. Все значения мер объединены в табл. 1.

Результаты показывают, что в настоящем примере низкая относительная чувствительность к параметрам (малые *RSS* и *RSM*) при подгонке к дикому типу (за счет насыщения) порождает неоднозначность предсказаний поведения мутантов.

Для повышения предсказательной способности модели дополнительно необходимы данные по экспрессии мутантов. Точную информацию о виде экспрессии РНК в мутанте из имеющихся изображений получить затруднительно, поэтому мы создаем искусственные паттерны с разной высотой пика PS4 в *Kr*-мутанте и используем их для подгонки. Если подгонка осуществляется одновременно к данным по двум генотипам, точность предсказаний становится выше, что отражается на значениях *RCM* (табл. 1). Значительно более высокие значения меры *RCM* говорят об отсутствии высоких корреляций между жесткими параметрами модели мутантов и остальными параметрами модели дикого типа, т.е. о большей надежности оценок.

На основании проведенного анализа мы приходим к выводу, что в силу нечувствительности полной модели проблемы с предсказаниями могут быть связаны с недостаточностью данных эксперимента, как о распределении экспрессии мРНК hb отдельно с двух промоторов, так и мутантных данных. Следовательно, маловероятно, что данная сильно упрощенная модель может обеспечить точное предсказание мутантов без дополнительных допущений.

Пример 2. Вложенные модели. При разработке модели [13] рассматривались разные комбинации транскрипционных факторов и разные виды функционала Ф, которые обеспечивают наилучшее соответствие данным эксперимента. Однако при этом важно знать, не приводит ли усложнение модели к переподгонке и вследствие этого к неправильным предсказаниям. Сначала исследуем, как добавление одного дополнительного транскрипционного фактора, Nubbin (Nub), в уравнение (8) меняет качество подгонки и предсказательные свойства модели, а затем с помощью нашего метода связываем предсказательные свойства с чувствительностью моделей. Репрессивное воздействие Nub на Kr позволяет установить постериорную (правую) границу пика PS4, что моделируется за счет включения в уравнение дополнительного регуляторного члена – θ_{Nub}[Nub]. Таким образом, в мо-

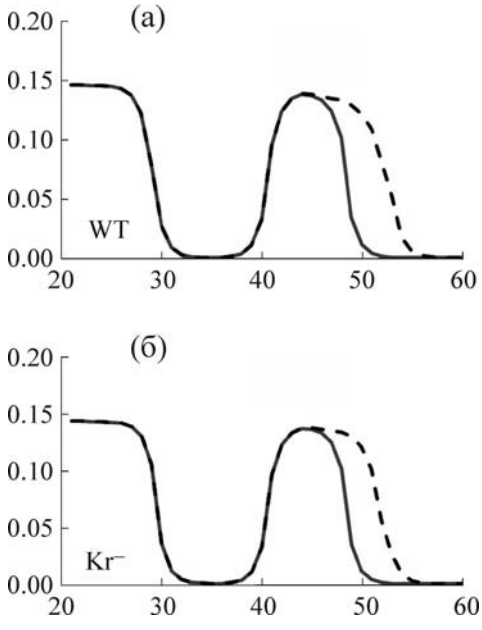

Рис. 5. Пример 2а. Варианты решения модели с Nub (сплошная линия) и без Nub (пунктирная линия): (а) – дикий тип (полная модель), (б) – *Kr*⁻-мутант (предсказываемая модель).

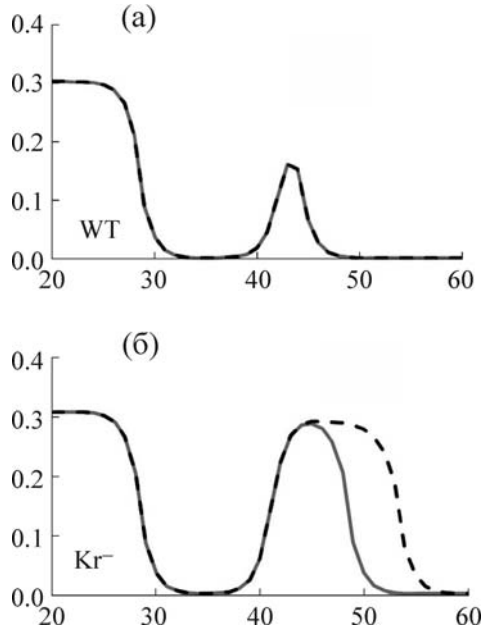

Рис. 6. Пример 2б. Варианты решения модели с Nub (сплошная линия) и без Nub (пунктирная линия): (а) – дикий тип (полная модель), (б) – *Kr*⁻-мутант (предсказываемая модель).

дели положение границы регулируется значением параметра θ_{Nub}, что позволяет улучшить подгонку к данным.

В то же время следует проверить, как усложнение модели влияет на ее предсказательные свойства. Например, если модель слабо чувствительна к параметру θ_{Nub}, то она правильно воспроизводит паттерны экспрессии в диком типе в широком диапазоне значений этого параметра, однако может оказаться, что не весь диапазон значений приводит к корректному предсказанию мутантов. Подгонка модели одновременно к данному дикому типу и мутантам позволяет сузить диапазон значений параметров, обеспечивающих верные предсказания.

Рассмотрим в качестве примера два характерных решения вышеприведенной модели (9) формирования пика PS4 на паттерне экспрессии промотора P2 гена *hb*. Оба решения демонстрируют хорошее соответствие данным с разной высотой пика. Демонстрируем как относительные меры чувствительности, вычисленные для моделей с Nub и без Nub, характеризуют изменение предсказательных свойств моделей за счет включения дополнительного транскрипционного фактора и связанного с ним параметра.

Пример 2а. Решение модели, глобально чувствительное к дополнительному параметру (представлено на рис. 5).

В данном примере решение локально вообще нечувствительно к Nub, т.е. частные производные решения по соответствующему параметру равны нулю и, следовательно, его доверительные интервалы бесконечны. Однако, несмотря на это, включение Nub в модель необходимо для правильного воспроизведения постериорной границы PS4, как показано на рис. 5а. При отсутствии Nub ($\theta_{Nub} = 0$) у паттернов как дикого типа, так и *Kr*-мутанта постериорная граница сдвигается вправо. В силу сигмоидного характера уравнения модели влияние Nub сказывается скачкообразно при определенном уровне его экспрессии, т.е. при значениях θ_{Nub}, превышающих некоторый порог. Особенность решения модели состоит в том, что при всех значениях параметра θ_{Nub} предсказываемый паттерн *Kr*-мутанта полностью совпадает с диким типом. Таким образом, подгоняемая модель дикого типа содержит информацию, достаточную для предсказания мутанта, что выражается в высоких значениях относительных мер чувствительности *RSM* (0,99 и 0,96) и *RSS* (>1) как для модели с включением Nub, так и более простой модели без Nub.

Пример 2б. Решение модели, глобально нечувствительное к дополнительным параметрам (представлено на рис. 6).

В этом примере модель P2 формирует пик PS4 меньшей высоты, т.е. не демонстрирует низкую чувствительность к своим параметрам

Таблица 2. Пример 2. Вложенные модели

	RSM	*RSS*	*RCM*
Пример 2а			
Модель с Nub	0,99	1,64	0,05
Модель без Nub	0,96	1,42	0,06
Пример 2б			
Модель с Nub	0,33	0,02	1,50E–05
Модель без Nub	0,00027	1,11E–05	0,005

Примечание. Меры вычисляются в области постериорной границы домена PS4 от 40 до 65% EL.

в целом. При этом решение дикого типа отличается тем, что не только локально нечувствительно к параметру θ_{Nub}, но и совсем от него не зависит. Таким образом, в данном примере включение Nub не очевидно является обязательным фактором. Решение модели без Nub хорошо описывает экспериментальные данные дикого типа, однако как видно из рис. 6б, в предсказываемом паттерне *Kr*⁻-мутанта ширина области PS4 превышает экспериментальную и сдвигается с увеличением значения θ_{Nub}. Понятно, что в такой ситуации подгонка модели только к данным дикого типа не может обеспечить точного предсказания мутанта, для которого необходимо знать значение параметра θ_{Nub}. Эта неопределенность отражается в значениях *RSM* ($\sim 10^{-4}$) и *RSS* ($\sim 10^{-5}$), очень близких к нулю. Для сравнения в табл. 2 приведены значительно более высокие значения обеих мер (0,33 и 0,02) для оценки предсказательной силы *Kr*⁻-мутанта в полной модели, которая за счет включения Nub потенциально обладает способностью более точных предсказаний.

Таким образом, результаты анализа показывают, что два приведенных примера принципиально различаются по своим предсказательным свойствам. Во втором случае зависимость решения от Nub не проявляется при подгонке, но очевидным образом определяет поведение модели в мутанте, в то время как в первом случае, напротив, присутствие в модели Nub меняет вид паттерна при подгонке к дикому типу, но никак не влияет на предсказания. Мы показали, что это различие является следствием чувствительности моделей к параметрам модели при обнулении θ_{Nub}. Низкая чувствительность во втором примере является причиной ненадежных оценок и неопределенности в предсказаниях. Наш анализ, таким образом, подтверждает необходимость усложнения модели даже в том случае, если это неочевидно по результатам подгонки.

ВЫВОДЫ

Биологические механизмы, ответственные за робастность системы к внешним возмущениям и внутреннему шуму, естественным образом моделируются при помощи насыщающих функций отклика. Математически это влечет низкую чувствительность или нечувствительность к некоторым параметрам модели в определенной области их значений. Нами рассматривается модель, в которой регуляция генов описывается насыщающей сигмоидной функцией и исследуются предсказательные свойства модели при изменении (в частности, обнулении) некоторых входных данных.

В качестве количественной характеристики предсказательной силы модели мы вводим две меры относительной чувствительности *RSM* и *RSS*, каждая из которых свидетельствует о том, достаточно ли точны оценки параметров модели, полученные подгонкой к данным, для точных предсказаний. При этом значения двух мер отражают различные возможные источники плохих предсказаний: (1) низкую чувствительность полной модели и (2) высокую чувствительность предсказываемой модели, требующую большой точности оценок ее параметров для предсказания правильного поведения системы.

Помимо низкой чувствительности модели существуют также и другие источники плохих предсказаний. Так, наличие сильных корреляций между обнуляемыми параметрами и теми, значения которых определяются в результате подгонки, может также быть препятствием для точных предсказаний. Мы вводим показатель *RCM*, отражающий этот источник неопределенности. Близкие к нулю значения хотя бы одной из мер свидетельствуют о низкой предсказательной силе при изменении/обнулении входных данных.

Метод оценки предсказательной силы применяется для анализа модели формирования паттерна экспрессии мРНК гена *hb* в эмбрионе дрозофилы с учетом модульной организации, опубликованной в [13]. Подгонка модели произведена к данным по экспрессии транскриптов *hb* в эмбрионе дикого типа, и проверяется способность модели воспроизводить правильные паттерны экспрессии в эмбрионах, нуль-мутантных по гену *Kr*. Было исследовано несколько решений модели, на которых продемонстрированы основные принципы применения нашего метода.

В приведенных примерах мы наблюдаем неопределенность в оценках параметров, приводящую к неоднозначным предсказаниям. Та-

кая ситуация возникает, например, если полная модель (дикого типа) достигает насыщения, т.е. существует множество решений модели, обеспечивающих равно хорошее согласие с данными, но предсказывающих совершенно разные паттерны экспрессии в мутанте. В таком случае для преодоления неоднозначности требуется дополнительная информация, без которой невозможны предсказания. Если же такие данные недоступны, то трудно ожидать надежных предсказаний. В случае если решение не является насыщенным, модель чувствительна к параметрам и их оценки более надежны, то модель способна к хорошим предсказаниям. Наш метод позволяет охарактеризовать решения модели с точки зрения чувствительности и выбрать среди вариантов моделей и их допустимых решений те, которые способны обеспечивать однозначные предсказания без привлечения дополнительных данных (которые могут быть недоступны).

Следует подчеркнуть, что, очевидно, наш анализ способен только выявить причины плохой предсказательности в силу неопределенности в оценках параметров, но не вследствие неадекватности модели.

Также показано, как метод позволяет определить допустимую степень детализации модели, не приводящую к переподгонке и недостоверным предсказаниям. Приведен пример того, как включение в модель дополнительных входных данных и параметров хоть и не влияет на качество подгонки, но в силу низкой чувствительности к параметрам меньшей модели сильно изменяет предсказания мутантов. Таким образом, руководствуясь значениями мер относительной чувствительности, делаем вывод о необходимости усложнения модели для верных предсказаний.

Наш метод явно учитывает свойство насыщения, присущее сигмоидной функции, однако применение метода отнюдь не ограничивается только этим случаем. Низкая чувствительность модели может являться следствием и другого математического представления нелинейности системы, например, за счет функции Хилла [18], экспоненты и др. Так что наш подход может быть естественным образом распространен и на такие модели.

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (гранты №№ 15-04-07800 и 15-04-06480) и Министерства образования и науки Российской Федерации (грант № 16.8549.2017/8.9).

СПИСОК ЛИТЕРАТУРЫ

1. M. Transtrum, B. Machta, K. Brown, et al., J. Chem. Phys. **143**, 010901 (2015).
2. А. Сальтелли и И. М. Соболь, Мат. моделирование **7**, 16 (1995).
3. J. Dresch, X. Liu, D. Arnosti, and A. Ay, BMC Syst. Biol. **4**, 142 (2010).
4. Z Zi, IET Syst. Biol. **5**, 336 (2011).
5. R. Gutenkunst, J. Waterfall, F. Casey, et al., PLoS Comput. Biol. **3** (10), 1871 (2007).
6. D. Bates and D. Watts, *Nonlinear Regression Analysis and its Applications* (J. Wiley, 1988).
7. M. Ashyraliyev, J. Jaeger, and J. Blom, BMC Syst. Biol. **2**, 83 (2008).
8. J. Jaeger, S. Surkova, M. Blagov, et al., Nature **430**, 368 (2004).
9. J. Jaeger, M. Blagov, D. Kosman, et al., Genetics **167**, 1721 (2004).
10. H. Janssens, S. Hou, J. Jaeger, et al., Nature Genetics **38**, 1159 (2006).
11. M. Manu, S. Surkova, A. Spirov, V. Gursky, et al., PLoS Biol. **7** (3): e1000049 (2009).
12. K. Kozlov, S. Surkova, E. Myasnikova, et al., PLoS Comput. Biol. **8** (8), e1002635 (2012).
13. A. Spirov, E. Myasnikova, and D. Holloway, J. Bioinform. Comput. Biol. **14**, 1641005 (2016).
14. E. Myasnikova and K. Kozlov, J. Bioinform. Comput. Biol. **12**, 1441002 (2014).
15. S. Surkova, E. Golubkova, M. Manu, et al., Dev. Biol. **376**, 99 (2013).
16. D. Clyde, M. Corado, X. Wu, et al., Nature **426**, 849 (2003).
17. K. Wotton, E. Jiménez-Guri, A. Crombach, et al., eLife, **4**, e04785 (2015).
18. J. Margolis, M. Borowsky, E. Steingrimsson, et al., Development **121**, 3067 (1995).
19. A. Spirov and D. Holloway, in *Evolutionary Computation in Gene Regulatory Network Research,* Ed. by H. Iba and N. Noman (John Wiley & Sons, Hoboken, NJ, USA, 2016), pp. 240–269.

Method for Estimating the Predictive Power in the Model of a Biological System with Low Sensitivity to the Parameters

E.M. Myasnikova* and A.V. Spirov**

**Peter the Great St. Petersburg Polytechnic University, ul. Polytechnicheskaya 29, St. Petersburg, 195251 Russia*

***Sechenov Institute of Evolutionary Physiology and Biochemistry, Russian Academy of Sciences, prosp. Toreza 44, St. Petersburg, 194223 Russia*

Low sensitivity of the model to perturbations of input data (the parameters) reflecting mathematically biological mechanisms of robustness may result in ambiguity of parameter estimation of the biological system model. We present a novel method for estimating the predictive power in the model of a biological system under the parameter estimation uncertainty. The predictions are understood as the model ability to correctly reproduce the system behavior with the altered inputs. The method is based on the analysis of relative sensitivity of the fitted model to stiff parameters of the predicted model. Application principles of our approach are demonstrated using the model for the formation of the pattern of mRNA expression of the *hb* gene in Drosophila embryo and its ability to predict the pattern of the *hb* gene in a null-mutant for *Kr* gene. A saturating sigmoid function is used for modeling of the system nonlinearity, it is the reason of low model sensitivity. Our method enables us to estimate the predictive power in the model and uncover the sources of poor predictions, as well as may give a clue in terms of predictions to a correct choice of the level of the model detail.

Keywords: mathematical model, biological system, sensitivity analysis, predictive power