Численные алгоритмы анализа чувствительности и сложности описания в задачах идентификации моделей математической иммунологии [$]

Г. А. Бочаров[@], Н. А. Медведева[◇]

Развитие эффективных вычислительных подходов к реализации задач системного анализа иммунных процессов в норме и при вирусных заболеваниях, с учетом блочной структуры организма, является единственным средством интеграции различных данных наблюдений и теоретических гипотез в единую математическую теорию иммунной системы. Модульный подход к построению уравнений моделей на основе балансных соотношений позволяет сформировать некоторое семейство возможных математических моделей. Проверка адекватности структуры уравнений данным наблюдений с целью выбора оптимальной модели является чрезвычайно проблемным и плохо исследованным этапом формирования количественных математических теорий (гипотез) иммунных процессов. Это связано с трудностями многократного решения обратных задач, численной реализации алгоритмов анализа идентифицируемости моделей, оценивания информационной сложности моделей. В данной работе рассматриваются ключевые элементы численной технологии построения оптимального описания динамики иммунных процессов на примере противовирусной реакции системы интерферона. В основе нашего подхода лежит

[$]Данная работа проводилась при поддержке Российского фонда фундаментальных исследований (грант РФФИ №05-01-00732).

[@]Институт вычислительной математики РАН

[◇]Московский государственный университет имени М.В.Ломоносова

использование принципа максимального правдоподобия для идентификации параметров модели, использование информационной матрицы Фишера для оценки степени неопределённости параметров, локальный и глобальный анализ чувствительности в рамках детерминистского и стохастического подходов, оценивание качества моделей на основе информационно-теоретических подходов.

1. Уравнения математической иммунологии

Одной из центральных задач математической иммунологии является разработка эффективной вычислительной методологии построения оптимальных, в смысле информативности, моделей динамики сложных систем. Это связано с решением задач усвоения данных, получаемых с помощью современных высокопроизводительных лабораторных и клинических методов, таких как, проточная цитофлурометрия, протеомика, геномика и др. В отличие от классической физики или химии, при моделировании живых систем используется феноменологический подход к конструированию уравнений моделей. При этом, модели сложных систем строятся из блоков описывающих элементарные процессы. Так модели популяционной динамики иммунных реакция строятся на основе балансных соотношений процессов рождения и гибели клеток и патогенов. Этот подход позволяет сформировать некоторое семейство возможных математических моделей, различающихся функциональными зависимостями, используемыми при описании одного и того же элементарного процесса, числом параметров, структурной сложностью. Анализ адекватности структуры моделей данным наблюдений с целью выбора наиболее информативной предполагает необходимость развития эффективных, адаптированных к конкретным типам уравнений, численных методов решения задач оценивания параметров моделей, исследования чувствительности и информационной сложности моделей. В данной работе излагается ключевые элементы единого подхода, на основе метода максимального правдоподобия, к построению оптимального описания иммунных процессов. В качестве конкрет-

ного объекта моделирования будет рассмотрена динамика реакции системы интерферона на уровне базового блока, описывающего активность дендритных клеток. В качестве средства математического описания мы рассматриваем модели на основе систем дифференциальных уравнений с запаздывающим аргументом:

$$\mathbf{y}'(t) \quad = \quad \mathbf{f}(\mathbf{y}(t), \mathbf{y}(t-\tau)), \mathbf{p}) \quad (\tau \geqslant 0), \qquad t \in [t_0, T], \tag{1}$$

где $\mathbf{y}(t, \mathbf{p}) \in \mathbb{R}^M$ и $\mathbf{p} \in \mathbb{R}^L$. Обозначим компоненты вектора параметров в уравнениях $\mathbf{p}$ через p_ℓ. Величина запаздывания, $\tau \geqslant 0$, является дополнительным параметром, анализ которого является более сложным, чем для обычных параметров p_ℓ. Для данного класса систем требуется задать начальные условия $[t_0 - \tau, t_0]$, в общем случае, также зависящие от параметров:

$$\mathbf{y}(t, \mathbf{p}) \quad := \quad \boldsymbol{\psi}_0(t, \mathbf{p}), \qquad t \in [t_0 - \tau, t_0]. \tag{2}$$

Решение модели в моменты времени наблюдений $t_j \in [t_0, T]$, $\mathbf{y}(t_j, \mathbf{p})$ должно соответствовать по выбранной мере данным измерений $\{\mathbf{y}_j\}$. Детальное изложение различных аспектов построения уравнений моделей в иммунологии рассматривается в работах [1, 2].

2. Математическая модель системы интерферона

В модели реакции системы интерферона в ответ на вирусную инфекцию [3] рассматривается популяционная динамика численностей *вирусов* $V(t)$, *молекул интерферона* $I(t)$, *неинфицированных клеток* $C(t)$ и *инфицированных клеток*, продуцирующих интерферон $C_V(t)$. Уравнения модели содержат описание кинетики процессов, изображенных на рис. 1.

Система уравнения модели содержит описание скорости изменения численности популяций вирусов, молекул интерферона и клеток с помощью обыкновенных дифференциальных уравнений (ОДУ) и дифференциальных уравнений с запаздывающим аргументом (ДУ-

ЗА) следующего вида:

$$\frac{dV}{dt}(t) = \frac{\rho_V C_V(t-\tau_V)}{1+I(t)/\theta} - d_V V(t), \tag{3a}$$

$$\frac{dI}{dt}(t) = \rho_I C_V(t-\tau_I) - d_I I(t), \tag{3b}$$

$$\frac{dC}{dt}(t) = -\sigma V(t) C(t) - d_C(t) C(t), \tag{3c}$$

$$\frac{dC_V}{dt}(t) = \sigma V(t) C(t) - d_{CV}(t) C_V(t). \tag{3d}$$

Одной из особенностей данной модели является использование закона Гомпертца, при описании продолжительности жизни клеток. С учетом этого, кинетика гибели клеток определяется двумя параметрами — начальной скоростью гибели и темпом увеличения данной скорости. Обоснованность данного описания будет исследована позже с помощью информационных критериев.

Начальные данные для задачи Коши $(t = 0)$ имеют вид: $V(0) =$

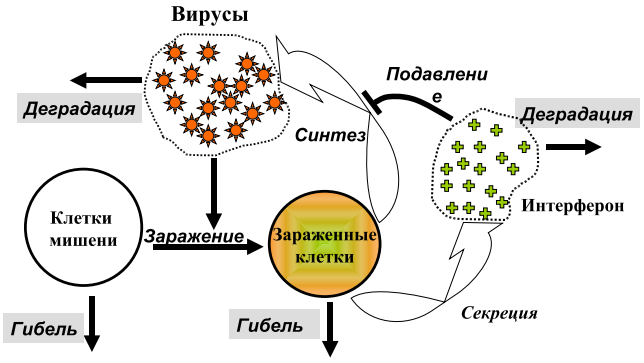

Рис. 1. Биологическая схема, лежащая в основе математической модели (3) синтеза интерферона антигенпрезентирующими клетками в ответ на вирусную инфекцию.

Параметр	Физический смысл	Размерность	Оптим. оценка
ρ_V	Скорость продукции вирусов одной клеткой	вир/час	1.1
ρ_I	Скорость продукции интерферона одной клеткой	пг/час	0.00091
θ	Порог 50% ингибирования продукции вирусов интерфероном	пг/мл	11.6
σ	Скорость инфицирования клеток	кл/вир/час	2.1×10^{-6}
τ_V	Задержка продукции вирусов клеткой	час	4.9
τ_I	Задержка продукции интерферона клеткой	час	4.5
d_{0C_V}	Начальная скорость гибели инфицированных клеток	1/час	0.1
k_{C_V}	Ускорение темпа гибели клеток инфицированных клеток	1/час	0.13
f_i	Начальная доля инфицированных клеток при кратности инфекции 1		0.0077
d_V	Скорость гибели вирусов	1/час	0.155
d_I	Скорость распада интреферона	1/час	0.012
d_{0C}	Начальная скорость гибели неинфицированных клеток	1/час	0.0055
k_C	Ускорение темпа гибели клеток неинфицированных клеток	1/час	0.089

Таблица 1. Параметры модели и их оценки полученные по методу максимального правдоподобия.

V_0, $I(0) = I_0$, $C(0) = C_0$, $C_V(t) = 0$, $t \in [-\max(\tau_V, \tau_I), 0]$. Параметры модели описаны в табл. 1. Для идентификации параметров моделей использовался подход на основе метода максимального правдоподобия.

3. Обратные задачи и критерий правдоподобия

Задача приближения модели к данным наблюдений связана, прежде всего, с выбором критерия близости или целевого функционала невязки. В идеале, этот выбор должен определяться ха-

рактером распределения и статистическими свойствами погрешностей измерений наблюдаемых переменных модели. Ранее, данный вопрос исследовался нами в работе [4]. В частности, путем анализа большого массива экспериментальных данных по динамике численности лимфоцитов при иммунном ответе было показано, что по критерию Колмогорова-Смирнова наиболее адекватной моделью ошибок является нормальное или лог-нормальное распределения. Функционал невязки $\Phi(\mathbf{p})$, $\Phi(\mathbf{p}) \geqslant 0$, зависит от данных наблюдений $\{t_j; y_j^i\}_{j=1}^N$ (для $i = 1, \cdots, M$) и значениями соответствующих переменных $\{y^i(t_j; \mathbf{p})\}_{j=1:N}^{i=1:M}$ решения модели $\mathbf{y}(t; \mathbf{p})$ (3), неявно зависящего от параметров уравнений. Тем самым, задача сводится к поиску таких значений параметров $\mathbf{p}^\star$ для (3), при которых решение $\{y^i(t_j; \mathbf{p}^\star)\}_{j=1:N}^{i=1:M}$, наилучшим образом описывает данные $\{y_j^i\}_{j=1:N}^{i=1:M}$:

$$\Phi(\mathbf{p}^\star) = \min_{\mathbf{p} \in \mathbb{R}_+^L} \Phi(\mathbf{p}). \tag{4}$$

Данная обратная задача, в общем случае, является некорректной в силу того, что с учетом ошибок измерений, статистически допустимым решениями являются все те значения параметров, при которых решение уклоняется от данных наблюдений не более чем на некоторую величину. Решение обратной задачи, с вычислительной точки зрения, сводится к поиску глобального минимума функционала $\Phi(\cdot)$.

Сделаем следующие предположения:

1) ошибки измерений в различные моменты времени независимы;

2) ошибки измерений распределены по нормальному закону

$$\mathbf{y}_j \sim \mathcal{N}(\mathbf{y}(t_j; \mathbf{p}), \Sigma_j),$$

где $\{\mathbf{y}(t_j; \mathbf{p})\}_{j=1}^N$ являются средними определяемыми моделью, а Σ_j является j-ой ковариационной матрицей (матрицей ошибок);

3) ошибки измерений компонент вектор-функции решения являются некоррелированными, т.е. ковариационная матрица является диагональной

$$\Sigma_j = \sigma^2 \left\{ \mathrm{diag}[\omega_1^{[j]}, \omega_2^{[j]}, \ldots, \omega_M^{[j]}] \right\}, \tag{5}$$

(где σ^2 — коэффициент вариации).

Для решения обратной задачи будем следовать принципу *максимального правдоподобия* (МП). Распределение вероятности наблюдения данных выборки объема N определяется произведением функций [5, 7]:

$$\left\{\mathcal{H}(\mathbf{y}_j;\mathbf{p}) = \frac{1}{\sqrt{(2\pi)^M \det \Sigma_j}} \exp\{-\frac{1}{2}[\mathbf{y}(t_j;\mathbf{p})-\mathbf{y}_j]^T \Sigma_j^{-1}[\mathbf{y}(t_j;\mathbf{p})-\mathbf{y}_j]\}\right\}_{j=1}^{N}. \tag{6}$$

Полная функция правдоподобия или вероятность получения данной выборки как функции вектора параметров модели $\mathbf{p}$ имеет вид

$$\mathcal{L}(\mathbf{p}) = \prod_{j=1}^{N} \mathcal{H}(\mathbf{y}_j;\mathbf{p}). \tag{7}$$

Вектор параметров $\mathbf{p}^\star$, при котором эта функция достигает максимума и является оценкой максимального правдоподобия.

Рассмотрим функционал взвешенных наименьших квадратов

$$\Phi_{\mathcal{W}LS}(\mathbf{p}) \equiv [\mathbf{y}(t_j;\mathbf{p}) - \mathbf{y}_j]^T \Sigma_j^{-1}[\mathbf{y}(t_j;\mathbf{p}) - \mathbf{y}_j]. \tag{8}$$

С учетом предположения 3)

$$\Phi_{\mathcal{W}LS}(\mathbf{p}) \equiv \sigma^{-2}\Phi_{\Omega LS}(\mathbf{p}), \tag{9}$$

где

$$\Phi_{\Omega LS}(\mathbf{p}) = \sum_j \|\mathrm{diag}^{-1}[\omega_1^{[j]}, \omega_2^{[j]}, \ldots, \omega_M^{[j]}][\mathbf{y}(t_j;\mathbf{p}) - \mathbf{y}_j]\|^2. \tag{10}$$

Логарифмическая функция максимального правдоподобия имеет вид

$$\begin{aligned} \ell n\ \mathcal{L}(\mathbf{p}) &= -\tfrac{1}{2}\Big\{NM\ \ell n\ (2\pi) + NM + 2\sum_{i,j}\ \ell n\ (\omega_i^{[j]})\Big\} \\ &\quad -\tfrac{1}{2}\{NM\ell n(\Phi_{\Omega LS}(\mathbf{p}) - NM\ell n(NM)\}. \end{aligned} \tag{11}$$

Нетрудно видеть, что максимизация функции правдоподобия $\mathcal{L}(\mathbf{p})$ эквивалентна минимизации $\Phi_{\Omega LS}(\mathbf{p})$ (или $\Phi_{\mathcal{W}LS}(\mathbf{p})$), при этом оценка МП дисперсии имеет вид

$$\begin{aligned}\sigma^{2\star} &= \frac{1}{NM}\sum_j \|\mathrm{diag}^{-1}[\omega_1^{[j]},\omega_2^{[j]},\ldots,\omega_M^{[j]}][\mathbf{y}(t_j,\mathbf{p}^\star)-\mathbf{y}_j]\|^2 \\ &= \frac{1}{NM}\Phi_{\Omega LS}(\mathbf{p}^\star). \end{aligned} \tag{12}$$

где $\mathbf{p}^\star$ обозначает оптимальную в смысле МП оценку параметров.

Задача поиска точечных оценок максимально правдоподобных параметров сводится к минимизации функционала наименьших квадратов. Корректный выбор конкретного вида функционала наименьших квадратов предполагает известным закон распределения ошибок измерений. В приложениях встречаются три варианта функционала невязки, соответствующих

- нормальному закону с дисперсией, не зависящей от моментов измерений и одинаковой для всех наблюдаемых переменных – обычный метод наименьших квадратов:

$$\Phi_{OLS}(\mathbf{p}) = \sum_{j=1}^{N}\sum_{i=1}^{M}\left[y^i(t_j;\mathbf{p})-y_j^i\right]^2 = \sum_{j=1}^{N}\|\mathbf{y}(t_j,\mathbf{p})-\mathbf{y}_j\|^2; \tag{13a}$$

- нормальному закону с дисперсией, не зависящей от моментов измерений, но различной для каждой переменных– взвешенный метод наименьших квадратов (при этом, может быть иметь место ситуация, когда относительная величина дисперсии одинакова для всех переменных):

$$\Phi_{\mathcal{W}LS}(\mathbf{p}) \equiv \sigma^{-2}\sum_{j=1}^{N}\sum_{i=1}^{M}\left\{\omega_i^{[j]}\left[y^i(t_j,\mathbf{p})-y_j^i\right]\right\}^2; \tag{13b}$$

- лог-нормальному закону (предполагается, что $y_j^i > 0$ и $y^i(t_j;\mathbf{p}) > 0$) с дисперсией не зависящей от моментов измерений и одинаковой (случай различных дисперсий приводит к взвешенному

варианту, аналогично предыдущему случаю) для всех наблюдаемых переменных – логарифмический метод наименьших квадратов:

$$\Phi_{\mathrm{LogLS}}(\mathbf{p}) = \sum_{j=1}^{N}\sum_{i=1}^{M}\left[\ \ln\left(y^i(t_j,\mathbf{p})\right) - \ \ln\left(y^i_j\right)\right]^2. \qquad (13c)$$

Нормальный и лог-нормальный законы распределения отражают, соответственно, арифметическую и геометрическую зависимость среднего от значений данных наблюдений. Качественно, различие между арифметической и геометрической нормальностью ошибок измерений, проявляется в том, в первом случае отклонения от ожидаемых средних значений в обе стороны (увеличения или уменьшения) вносят одинаковый вклад в значение функционала, если их абсолютные значения равны, а во втором варианте, только, если их относительные значения равны. Последнее вариант предполагает, что если масштаб переменных модели сильно варьирует, то следует использовать логарифмический метод наименьших квадратов. Для идентификации параметров модели 3 мы использовали логарифмичекий метод наименьших квадратов, с учетом того, что переменные модели существенно различаются по своим абсолютным значениям. Часть фундаментальных параметров модели (последние 4 в табл. 1, в частности, скорости деградации вирусов, интерферона и неинфицированных клеток, определялась по независимым экспериментальным данным. Оптимальным оценкам параметров табл. 1 соответствует решение модели приведённое на рис. 2. Заметим, что часть переменных модели ($V(t)$ и $I(t)$) являются непосредственно наблюдаемыми, а оставшиеся должны быть преобразованы в наблюдаемые некоторым нелинейным образом

$$\%\text{Live pDC}(t) = \frac{C(t)+C_V(t)}{C_0}, \qquad (14a)$$

$$\%\text{Infected pDC}(t) = \frac{C_V(t)}{C(t)+C_V(t)}. \qquad (14b)$$

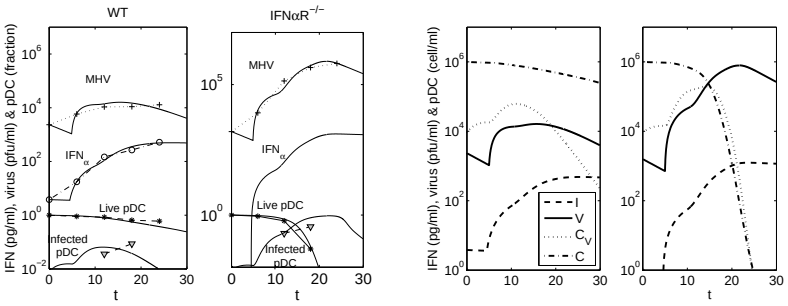

Рис. 2. Данные наблюдений и решения модели, соответствующие минимуму логарифмического функционала наименьших квадратов. Два левых графика описывают кинетику экспериментально наблюдаемых процессов в случае инфекции нормальных дендритных клеток и клеток, у которых отсутствует рецептор к интерферону. Два правых графика характеризуют динамику переменных модели, соответственно, при наличии ингибирующего эффекта интерферона на продукцию вирусов, и в случае, когда его нет.

4. Уравнения чувствительности и информационная матрица Фишера

4.1. Чувствительность по параметрам и запаздыванию. В основе анализа чувствительности моделей к вариациям параметров лежит исследование элементарных коэффициентов чувствительности, являющихся частными производными от компонент вектор-функции решения модели по параметрам. Предполагая гладкость решения $\mathbf{y}(t;\mathbf{p})$ по вектору параметров $\mathbf{p}$, рассмотрим разложение

$$\mathbf{y}(t;\mathbf{p}+\delta\mathbf{p}) = \mathbf{y}(t;\mathbf{p}) + \mathbf{S}(t,\mathbf{p})\delta\mathbf{p} + O(\|\delta\mathbf{p}\|^2).$$

Матрица $\mathbf{S}(t) \equiv \mathbf{S}(t,\mathbf{p})$ является $M \times L$ *матрицей коэффициентов чувствительности*, i-я строка которой имеет вид $\mathbf{s}_i(t;\mathbf{p}) = \left[\dfrac{\partial y^i(t,\mathbf{p})}{\partial p_1}, \dfrac{\partial y^i(t;\mathbf{p})}{\partial p_2}, \ldots, \dfrac{\partial y^i(t;\mathbf{p})}{\partial p_L}\right]^T$. Используя обозначение $\left\{\dfrac{\partial}{\partial \mathbf{p}}\right\}^T$, матрицу коэффициентов чувствительности $\mathbf{S}(t;\mathbf{p})$ можно записать

в виде

$$\mathbf{S}(t;\mathbf{p}) \equiv \left\{\frac{\partial}{\partial \mathbf{p}}\right\}^{T} \mathbf{y}(t;\mathbf{p}) \in \mathbb{R}^{M\times L}. \tag{15}$$

Таким образом, матрица $\mathbf{S}(t;\mathbf{p})$ характеризует локальную чувствительность решения модели $\mathbf{y}(t,\mathbf{p})$ к *малым* изменениям компонент параметра $\mathbf{p}$; при этом, строка $\mathbf{s}_i(t;\mathbf{p})$ соответствует первым производным i-ой компоненты $\mathbf{y}_i(t;\mathbf{p})$ по параметрам p_ℓ ($\ell \in \{1,2\ldots,L\}$).

Особенностью ДУЗА является разрывность первых производных решения по времени. Если начальная функция имеет разрыв первого рода в точке t_0, то соотвествующее решение имеет разрывы производных в моменты $t \in \bigcup\limits_{n\in\mathbb{N}} n\tau$. При этом, имеет место увеличение гладкости решений с ростом t. Воспользуемся обозначениями

$$\mathbf{y} := \mathbf{y}(t;\mathbf{p}), \quad \mathbf{y}_\tau := \mathbf{y}(t-\tau;\mathbf{p}), \quad \mathbf{y'}_\tau := \mathbf{y}'(t-\tau;\mathbf{p}). \tag{16}$$

Для системы (1) зависящие от времени коэффициенты чувствительности решения по компонентам вектора параметров $\mathbf{p}$ удовлетворяют на отрезках $t \in \bigcup\limits_{n\in\mathbb{N}} [t_0+n\tau, t_0+(n+1)\tau]$ следующей системе уравнений чувствительности

$$\begin{aligned}\mathbf{S}'(t) \quad &= \frac{\partial}{\partial \mathbf{y}}\mathbf{f}(\mathbf{y},\mathbf{y}_\tau;\mathbf{p})\mathbf{S}(t) + \frac{\partial}{\partial \mathbf{y}_\tau}\mathbf{f}(\mathbf{y},\mathbf{y}_\tau;\mathbf{p})\mathbf{S}(t-\tau)\\ &\quad + \frac{\partial}{\partial \mathbf{p}}\mathbf{f}(\mathbf{y},\mathbf{y}_\tau;\mathbf{p}).\end{aligned} \tag{17}$$

На рис. 3 приведены графики поведения элементарных функций (коэффициентов) чувствительности решения модели 3 для нормальных клеток и клеток, не имеющих рецепторы к интерферону, по параметрам ρ_I и ρ_V, характеризующим скорости продукции одной клеткой интерферона и вирусных частиц.

Производная $\mathbf{s}_\tau(t) \equiv \mathbf{s}_\tau(t,\mathbf{p})$ вектор-функции решения $\mathbf{y}(t;\mathbf{p})$ по параметру запаздывания τ ($\mathbf{s}_\tau(t,\mathbf{p}) = \dfrac{\partial}{\partial\tau}\mathbf{y}(t;\mathbf{p}) \in \mathbb{R}^M$) удовлетворяет уравнениям с запаздывающим аргументом нейтрального типа

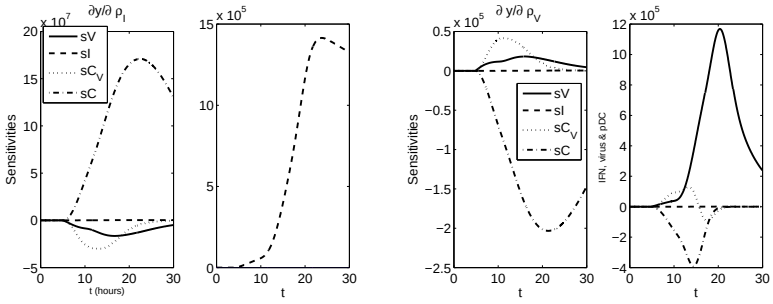

Рис. 3. Динамика производных решения модели по параметрам — коэффициенты чувствительности. Два левых графика описывают чувствительность решений к изменению скорости синтеза интерферона в случае нормальных дендритных клеток и клеток, у которых отсутствует рецептор к интерферону, соответственно. Аналогично, два правых графика характеризуют чувствительность решений модели к вариации скорости синтеза вирусов при наличии ингибирующего эффекта интерферона на продукцию вирусов, и в случае, когда его нет.

следующего вида:

$$\begin{aligned}\mathbf{s}_\tau'(t) &= \frac{\partial}{\partial \mathbf{y}}\mathbf{f}(\mathbf{y},\mathbf{y}_\tau,\mathbf{y}_\tau';\mathbf{p})\mathbf{s}_\tau(t)+\frac{\partial}{\partial \mathbf{y}_\tau}\mathbf{f}(\mathbf{y},\mathbf{y}_\tau,\mathbf{y}_\tau';\mathbf{p})\mathbf{s}_\tau(t-\tau)\\ &\quad -\frac{\partial}{\partial \mathbf{y}_\tau}\mathbf{f}(\mathbf{y},\mathbf{y}_\tau,\mathbf{y}_\tau';\mathbf{p})\mathbf{y}'(t-\tau)+\frac{\partial}{\partial \tau}\mathbf{f}(\mathbf{y},\mathbf{y}_\tau,\mathbf{y}_\tau';\mathbf{p}). \qquad (18)\end{aligned}$$

Численное решение таких уравнению представляет более сложную задачу, т.к. увеличения гладкости решений с ростом t в случае уравнений нейтрального типа не имеет места [6].

4.2. Информационная матрица Фишера. При оценивании точности решения обратных задач фундаментальная роль принадлежит матрице Фишера, которая характеризует количество информации о параметрах модели содержащейся в конкретной выборке данных наблюдений [8, 9]. Информационная матрица позволяет определить предельную точность, с которой можно оценить параметра модели. Информационная матрица Фишера $\mathbf{F}$, определяется как

математическое ожидание матрицы вторых производных функции максимального правдоподобия по элементам вектора параметров $\mathbf{p}$

$$\mathbf{F}(\mathbf{p}^\star, \tau^\star) \equiv \mathrm{E}\Big\{ \frac{\partial^2}{\partial \mathbf{p}^2} \mathcal{L}(\mathbf{p}, \tau) \Big\}_{\mathbf{p}^\star, \tau^\star}. \tag{19}$$

Используя решение уравнений чувствительности $\mathbf{S}, \mathbf{s}_\tau$, для множества времен данных наблюдений $t_n{}_{n=1}^N$ построим расширенную матрицу чувствительности $\mathbf{Z}$, строки которой содержат коэффициенты чувствительности по параметрам для всех моментов (выборки) измерений

$$\mathbf{Z} := \begin{bmatrix} \mathbf{S}(t_1; \mathbf{p}^\star, \tau^\star) & \mathbf{s}_\tau(t_1; \mathbf{p}^\star, \tau^\star) \\ \mathbf{S}(t_2; \mathbf{p}^\star, \tau^\star) & \mathbf{s}_\tau(t_2; \mathbf{p}^\star, \tau^\star) \\ \vdots & \vdots \\ \mathbf{S}(t_N; \mathbf{p}^\star, \tau^\star) & \mathbf{s}_\tau(t_N; \mathbf{p}^\star, \tau^\star) \end{bmatrix} \tag{20}$$

В случае, когда ошибки наблюдений распределены по нормальному закону, то для матрицы Фишера справедливо следующее представление через выборочную матрицу чувствительности $\mathbf{Z}$:

$$\mathbf{F} := \mathbf{Z}^\mathrm{T} \Sigma_\mathbf{y} \mathbf{Z}. \tag{21}$$

Согласно неравенству информации Крамера–Рао нижняя оценка элементов матрицы рассеяния параметров $\Sigma_\mathbf{p}$ (ковариационной матрицы параметров) от истинных значений определяется элементами матрицы, обратной информационной матрице Фишера:

$$\Sigma_\mathbf{p} \geqslant \mathbf{F}^{-1}. \tag{22}$$

Результаты расчета матрицы Фишера для анализа предельной оценки точности важнейших параметров модели интерферона (за исключением запаздываний и начальной доли инфицированных клеток) приведены в табл. 2. Полученные оценки снизу для стандартных отклонений оценок параметров показывают, что по имеющемуся массиву данных измерений параметры, модели нельзя определить точнее, чем 50% от их номинальных значений.

5. Анализ идентифицируемости параметров

Выборочная матрица чувствительности позволяет ранжировать параметры по их вкладу в измеренные значения наблюдений. Поскольку масштаб переменных модели и параметров сильно варьирует, мы провели анализ перемасштабированной выборочной матрицы чувствительности, составленной из блоков

$$\mathbf{S}_{\log}(t;\mathbf{p}) \equiv \left\{\frac{\partial}{\partial \log(\mathbf{p})}\right\}^{T} \log(\mathbf{y}(t;\mathbf{p})) \in \mathbb{R}^{M\times L}. \tag{23}$$

Алгоритм анализа, приведённый в [10] для исследования задач химической кинетики, аналогичен методу линейной пошаговой регрессии: факторы последовательно включаются в регрессионную модель в порядке убывания корреляционной связи с откликом; при этом, остаточная дисперсия, показывающая степень расхождения между данными и прогнозируемыми моделью значениями отклика, может быть неприемлемо велика, для преодоления чего и пополняется модель. Схема численной реализации алгоритма можно записать в виде

S1. Вычислить сумму квадратов элементов для каждого столбца $\mathbf{Z}$;

S2. Выбрать столбец $\mathbf{Z}_l$ с максимальной суммой, $1 \leqslant l \leqslant L$;

Таблица 2. Предельная оценка точности параметров на основе информационной матицы Фишера.

Параметр	Оптимальная оценка	Нижняя оценка дисперсии
ρ_V	1.1	0.52
ρ_I	0.00091	0.0004
θ	11.6	6.5
σ	2.1×10^{-6}	8.9×10^{-7}
d_{0C_V}	0.1	0.1
k_{C_V}	0.13	0.055

S3. Вычислить коэффициенты линейной регрессии столбцов матрицы $\mathbf{Z}$ относительно столбцов расширенной матрицы $\mathbf{P}$, составленной следующим образом $\mathbf{P} = [\mathbf{P}, \mathbf{Z}_l]$ (при первой итерации $\mathbf{P} = \mathbf{Z}_l$, где l номер выбранного столбца) по методу наименьших квадратов и осуществить прогноз:

$$\widehat{\mathbf{Z}} = \mathbf{P}(\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}\mathbf{Z}; \tag{24}$$

S4. Вычислить матрицу невязки (ошибку регрессионного прогноза) $\mathbf{R} = \mathbf{Z} - \widehat{\mathbf{Z}}$;

S5. Переопределить анализируемую матрицу $\mathbf{Z} := \mathbf{R}$;

S6. Если число столбцов матрицы $\mathbf{P}$ меньше L, то повторить шаги 1–5.

Применив данный алгоритм для анализа для матрицы чувствительности, параметры модели можно упорядочить по их относительному вкладу в наблюдаемые переменные модели следующим образом

$$k_{C_V} > \sigma > \rho_V > \rho_I > \theta > d_{0C_V}.$$

Результаты анализа позволяют более точно оценить вклад параметров в динамику наблюдаемых переменных, с учетом особенностей эксперимента.

6. Информационные критерии сложности моделей

Поскольку для описания количественных закономерностей одного и того же динамического процесса в иммунологии можно предложить несколько различных моделей, важнейшей задачей анализа является выбор наиболее адекватной модели. Величина функционала невязки $\Phi(\mathbf{p}^\star)$ в точке минимума является простейшей характеристикой близости конкретной модели $\mathbf{y}(t_j; \mathbf{p})$ к данным $\mathbf{y}_j$. В общем случае, данный критерий не является обоснованным, поскольку модели более сложной структуры, например, с большим числом параметров, могут более точно приблизить данные. При этом, поскольку информационное содержание массива данных остается

неизменным, естественно ожидать, что начиная с некоторого уровня сложности, точность оценивания параметров будет уменьшаться. Это, в свою очередь, ухудшит качество прогноза на основе модели. Следуя классической концепции информационного расстояния Кульбака-Лейблера и интерпретации математической модели как предполагаемой функции плотности распределения вероятности наблюдений, целью задачи идентификации является отыскание модели, минимизирующей информационное расстояние до истинной системы [9]. Подход к идентификации связанный с максимизацией функции правдоподобия позволяет оценить среднее информационное расстояние между данной моделью и истинной системой. Широко известен информационный критерий Акаике ранжирования моделей [11]:

$$\mu = -2\,\ell n\,\mathcal{L}(\widehat{\mathbf{p}}) + 2(L+1), \tag{25a}$$

$$\mu_c = -2\,\ell n\,\mathcal{L}(\widehat{\mathbf{p}}) + 2(L+1) + \frac{2(L+1)(L+2)}{\nu - L - 2}, \tag{25b}$$

$$\nu = NM.$$

Модели с меньшей величиной критерия находятся ближе к истинной системе по информационной мере Кульбака-Лейблера. Вопросы приложений данного критерия в задачах математической иммунологии исследовались нами, в частности в [7].

Наряду с критериями, в основе которых лежит оценивание информационного уклонения от идеальной модели данных наблюдений, перспективным представляется подход к ранжированию моделей по критерию «длины описания» [14], связанный с именем Риссанена. В основе данного критерия лежит концепция Колмогоровской сложности алгоритмов описания данных. В рамках данного подхода наилучшей является модель, которая допускает наиболее компактное описание данных, т.е. наиболее сильное сжатие информации в данных [9].

$$\Delta_{MDL} = -\ell n\mathcal{L}(\widehat{\mathbf{p}}) + \frac{L}{2}\log(\frac{\nu}{2\pi}) + \log(\int_{\Omega}\sqrt{\det[\mathbf{F}(\mathbf{p})]}\,d\mathbf{p}) \tag{26}$$

Реализация алгоритма вычисления длины описания для многопараметрических моделей является вычислительно трудоёмкой задачей,

т.к. связана с расчетом многомерных интегралов, зависящих от детерминанта матрицы Фишера, в свою очередь, определяемой решением системы уравнений чувствительности модели по параметрам.

Для иллюстрации эффективности данных критериев, рассмотрим простейшую задачу кинетики персистенции клеток в норме, т.е. при отсутствии возмущений в виде инфекции. Наиболее распространённым способом описания кинетики является экспоненциальная модель (обозначим её индексом E)

$$\frac{d}{dt}C(t) = -d_C \cdot C(t), \quad C(0) = C_0. \tag{27}$$

В то же время, многие задачи биологии продолжительности жизни моделируются с помощью закона Гомпертца (обозначим G), задаваемого системой уравнений для гибели клеток и изменения параметра скорости гибели:

$$\frac{d}{dt}C(t) = -d_C(t) \cdot C(t), \tag{28a}$$

$$\frac{d}{dt}d_C(t) = k_C \cdot d_C(t). \tag{28b}$$

На рис. 4 представлены экспериментальные данные и решения модели соответствующие оптимальным оценкам параметров этих двух моделей. Величины функционалов невязки в точке минимума, и соответствующих критериев Акаике и длины описания для данных моделей имеют следующие значения:

$$\Phi_E = 1012,\ \Phi_G = 1012,\ \mu_E = 38,\ \mu_G = 30,\ \Delta_E = 22.1,\ \Delta_G = 19.5.$$

Таким образом, сочетание критериев близости к данным наблюдений, с информационными критериями, позволяет однозначно идентифицировать модель Гомпертца, как наиболее адекватный закон описания процесса гибели клеток в данной системе.

7. Глобальный анализ чувствительности

Изложенные выше методы анализа чувствительности моделей являлись локальными. В реальных приложениях важным являет-

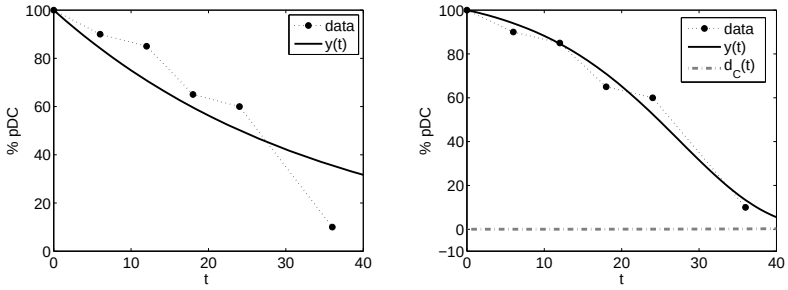

Рис. 4. Данные наблюдений и решения моделей экспоненциальной гибели клеток (слева) и гибели по закону Гомпертца (справа).

ся исследование вариаций глобального поведения модели при одновременном изменении параметров. Глобальный анализ чувствительности многопараметрических моделей можно провести, используя совокупность двух приемов: выборки на основе Латинского гиперкуба и рангового критерия оценки параметров [12, 13]. Результатом такого анализа является ранжирование параметров модели по степени их влияния на переменные модели или функционалы от таковых. Способ выборки по схеме Латинского гиперкуба является разновидностью метода Монте Карло. Каждый анализируемый параметр модели трактуется как случайная величина. Для всех параметров определяется функция распределения, область значения делится особым способом, и затем, параметр выбирается случайным образом. С помощью такой выборки можно получить множество сочетаний значений параметров, таких, что выборочное значение каждого параметра используется только один раз. Далее, модель просчитывается для всех наборов параметров, и полученные значения переменных модели используются для анализа чувствительности (корреляции) каждой полученной переменной и каждого параметра с применением рангового критерия.

7.1. Построение выборки на основе Латинского гиперкуба. Следуя [12, 13], для каждого исследуемого параметра из вектора параметров модели $\mathbf{p}$ необходимо определить область значений и

функцию плотности распределения. Далее, выбирается число испытаний N — количество наборов параметров и соответственно число прогонок модели. Строгие критерии для выбора этого числа не существуют, однако опытным путем ранее было получено, что $N > \frac{3}{4}L$. Соответственно, область значений каждого из L параметров делится на N неперекрывающихся, равновероятных отрезков. Далее, составляется таблица выборки на основе Латинского гиперкуба размером $N \times L$ по следующему правилу: на каждом из N интервалов случайным образом, но в соответствии с законом распределения, выбирается значение каждого параметра, после чего, снова случайным образом составляются пары из интервалов для первого и второго параметров, далее из этих пар опять случайным образом получают тройки с третьим параметром и так далее, пока не исчерпаны все параметры. В результате, получаем некоторую матрицу размером $N \times L$, строки которой содержат случайный вектор параметров $\mathbf{p}$, а каждый столбец содержит случайные реализации каждого параметра p_ℓ. Последним этапом реализации алгоритма реализации случайной выборки является численный расчет решений модели для всех векторов $\mathbf{p}$ из построенной выборки параметров, с целью получения набора значений исследуемых переменных модели.

Для исследуемой модели системы интерферона область неопределенности значений параметров задавалась в следующем виде: $\frac{1}{2}p_\ell^* \leqslant p_\ell \leqslant 2p_\ell^*$ и предполагалось, что параметры распределены по треугольному закону с пиком в точке $\frac{p_\ell^{max}+p_\ell^{min}}{2}$. В ходе численного эксперимента N равнялось 500. В качестве исследуемой переменной (целевой функционал) бралось суммарное количество интерферона $I_t \equiv \int_0^T I(t)dt$.

7.2. Ранговый критерий в оценке значимости параметров.. Чтобы ранжировать параметры по степени их влияния на целевой функционал модели, будем использовать следующий статистический критерий. Для множества векторов параметров $\mathbf{p}$ и значений зависимой переменной I_t, полученных в ходе реализации выборки на основе Латинского гиперкуба, составим матрицу $N \times (L+1)$ с данными расчетов по модели $X = \{X_l\}_{l=1}^{l=L+1}$. Далее, вычислим корреляционную матрицу C, элементы которой C_{ij} являются коэффи-

циентами корреляции между переменными столбцов X_i и X_j:

$$C_{ij} = \frac{(L+1)\sum X_i^k X_j^k - \sum X_i^k \sum X_j^k}{\sqrt{(L+1)\sum (X_i^k)^2 - (\sum X_i^k)^2}\sqrt{(L+1)\sum (X_j^k)^2 - (\sum X_j^k)^2}} \quad (29)$$

Поскольку нас интересует корреляция функционала I_t и параметров моедли, определим частичный коэффициент корреляции между I_t и p_ℓ по формуле (C^{-1} – обратная матрица):

$$\sigma_{p_l I_t} = -C_{l(L+1)}/(C_{ll}C_{(L+1)(L+1)})^{\frac{1}{2}}. \quad (30)$$

По результатам анализа данных коэффициентов параметры модели можно упорядочить по их степени их влияния на изменчивость I_t.

Применение изложенного алгоритма к модели системы интерферона позволяет оценить неопределённость в значении функции от решения модели и случайными значениями параметров. Результаты расчетов для для некоторых параметров приведены на рис. 5-6. Сплошные линии, представленные на графиках соответствуют уравнениям регресссии. Интересно, что зависимость I_t от параметра запаздывания τ_V может быть неплохо описана экспоненциальной кривой, в то время как по остальным параметрам такой явной зависимости нет. Количественно, анализ корреляции приводит к следующим оценкам коэффициентов ранговой корреляции между целевой функцией и параметрами модели, представленными в табл. 3.

Величина коэффициента корреляции характеризует степень влияния неопределенности значения параметра на неточность в оценке целевой функции. Для рассматриваемой модели можно сделать

Таблица 3. Выборочные коэффициенты ранговой корреляции между I_t и параметрами модели интерферона.

	I_t		I_t		I_t
ρ_I	0.3917	σ_V	0.0684	τ_I	−0.0464
ρ_V	−0.0007	d_{0c}	0.0151	τ_ν	−0.0947
θ	0.0286	k_c	0.0266	f_i	0.3154

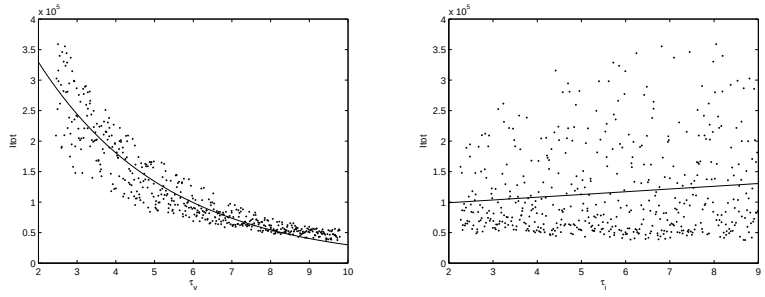

Рис. 5. Результаты численных экспериментов по исследованию чувствительности модели интерферона (зависимая переменная – I_t) к параметрам модели с построением случайной выборки по схеме Латинского гиперкуба. Слева: корреляция I_t с величиной запаздывания τ_V, уравнение регрессии имеет вид $I_t = 6 \times 10^5 \exp^{-0.3 \cdot \tau_V}$. Справа: корреляция I_t с величиной запаздывания τ_I, уравнение регрессии имеет вид $I_t = 90000 + 4500 \cdot \tau_V$.

вывод о том, что наиболее значимыми с точки зрения глобальной корреляции суммарного количества интреферона произведенного в ходе инфекции является скорость синтеза интерферона ρ_I и число первично инфицированных клеток f_i.

8. Заключение: матричный анализ в задачах идентификации

Оценка параметров и идентификация оптимальных моделей является важнейшей прикладной задачей моделирования в биологии. Для её успешного решения требуется развитие эффективных методов решения обратных задач, оценивания параметрической идентифицируемости моделей и информационной сложности моделей. В данной работе представлен набор базовых алгоритмов анализа параметрических моделей в контексте данных наблюдений. Развиваемая нами вычислительная технология моделирования в иммунологии может быть использована для изучения гораздо более широкого спектра прикладных задач математической биологии. Вопросы

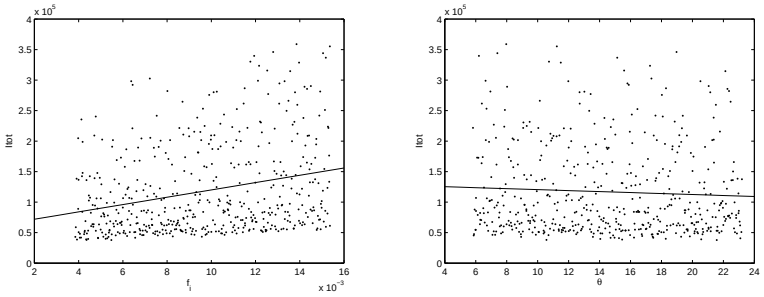

Рис. 6. Результаты численных экспериментов по исследованию чувствительности модели интерферона (зависимая переменная — I_t) к параметрам модели с построением случайной выборки по схеме Латинского гиперкуба. Слева: зависимость I_t от доли первично инфицированных клеток f_i, уравнение регрессии имеет вид $I_t = 60000 + 6 \times 10^6 \cdot f_i$. Справа: зависимость I_t от параметров ингибирования синтеза вирусов θ, уравнение регрессии имеет вид $I_t = 128500 - 803 \cdot \theta$.

эффективной численной реализации соответствующих алгоритмов базируются на широким применением методов матричного анализа и линейной алгебры [15, 16] и требуют дальнейших исследований.

Список литературы

[1] Бочаров Г.А., Марчук Г.И. Прикладные проблемы математического моделирования в иммунологии. *ЖВМиМФ.* // 2000. **40**. 1905–1920.

[2] Andrew S.M., Baker C.T.H., Bocharov G.A. Rival approaches to mathematical modelling in immunology. // *J. Comput. Appl. Math.* 2007. **205**. 669–686.

[3] G. Bocharov, L. Cervantes-Barragan, R. Zust, K. Eriksson, V. Thiel, B. Ludewig. Mathematical modeling of the antiviral type I interferon response. // In: Proceedings of the FOSBE 2007 Eds. F. Allgower and M.Reuss. Fraunhofer IRB Verlag. 2007. 325–330.

[4] L.K. Babadzanjanz, A.A. Voitylov, P. Krebs, B. Ludewig, D.R. Sarkissian, G.A. Bocharov. On primary statistical data processing of experimental measurements of lymphocytes using C57BL/6 mouse line. // В сб. «Устойчивость и процессы управления». Международной конференции посвященной 75-летию В.И. Зубова, под ред. Д.А. Овсянникова и Л.А.Петросяна. Санкт-Петербургский государственный Университет. 2005. Т. 2. С. 1227-1236.

[5] C.T.H. Baker, G.A. Bocharov, C.A.H. Paul and F.A. Rihan. Computational modelling with functional differential equations: identification, selection and sensitivity. // *Applied Numerical Mathematics*. 2005. **53**. 107–129.

[6] Christopher T.H. Baker and Gennady A. Bocharov. Computational aspects of time-lag models of Marchuk type that arise in immunology. // *Russ. J. Numer. Anal. Math. Modelling*. 2005. **20**. 247–262.

[7] C.T.H. Baker, G.A. Bocharov, J.M. Ford, P.M. Lumb, S.J. Norton, C.A.H. Paul, T. Junt, P. Krebs, B. Ludewig. Computational Approaches to Parameter Estimation and Model Selection in Immunology. // *J. Comput. Appl. Math*. 2005. **184**. 50–76.

[8] Теребиж В.Ю. Введение в статистическую теорию обратных задач. – М.: ФИЗМАТЛИТ. 2005.

[9] Льюнг Л. Идентификация систем. Теория для пользователя. – М.: Наука. 1991.

[10] K. Zhen Yao, Benjamin M. Shaw, Bo Kou, Kim B. McAuley, D. W. Bacon. Modeling Ethylene/Butene Copolymerization with Multi-site Catalysts: Parameter Estimability and Experimental Design. // *Polymer Reaction Engineering*. 2003. **11**. pp. 563–588.

[11] Burnham, K.P., Anderson, D.R. Model Selection and Multimodel Inference - a practical information-theoretic approach, 2nd ed. — Springer-Verlag, New York. 3rd printing. 2004.

[12] Ronald L. Iman, Jon C. Helton. An Investigation of Uncertainty and Sensitivity Analysis Techniques for Computer Models. // *Risk Analysis* 1988. **8** (1). 71–90.

[13] S. M. Blower. H. Dowlatabadi. Sensitivity and Uncertainty Analysis of Complex Models of Disease Transmission: An HIV Model, as an Example. // *International Statistical Review / Revue Internationale de Statistique*. 1994. **62**. pp. 229–243.

[14] Hanson A.J., Fu P.C-W. Applications of MDL to selected families of models. In: Advances in Minimum Description Lenght Theory and Applications. Ed. by P.D. Grünwald, I.J. Myung, M.A. Pitt. The MIT Press. Cambridge MA. 195–150. 2005.

[15] Воеводин В.В. Численные методы линейной алгебры (теория и алгоритмы). –М.: Наука. 1966.

[16] Дж. Голуб., Ч. Ван Лоун. Матричные вычисления. –М.: Мир. 1999.