УДК 004.8+519.7

К.С. ТЕПЛИНСКИЙ, асп., ДонНТУ, Красноармейск

ГИБРИДНАЯ ГЕНЕТИЧЕСКАЯ ИДЕНТИФИКАЦИЯ ПАРАМЕТРОВ НЕЛИНЕЙНЫХ БИОЛОГИЧЕСКИХ ДИНАМИЧЕСКИХ СИСТЕМ

Проведено исследование разработанного гибридного генетического алгоритма оптимизации параметров сложной нелинейной биологической модели трёхступенчатого биохимического метаболизма. Дополнительно описаны необходимые модификации генетического алгоритма, которые учитывают особенности биологических моделей. Ил.: 2. Табл.: 1. Библиогр.: 14 назв.

Ключевые слова: динамические системы, идентификация параметров, гибридный гнетический алгоритм.

Постановка проблемы. Разработка эффективных подходов к моделированию является важной задачей в системной биологии, которая обеспечивает новые средства анализа получаемых данных и базируется на глубоком понимании языка клеток и организмов. Впоследствии эти подходы могут стать базой для систематических стратегий получения ключевых результатов в медицине, фармацевтической и биотехнологической индустриях. Например, подходы, основанные на построении моделей, могут предоставлять необходимую инфраструктуру, способствующую производству медикаментов, принимая во внимание влияние этих медикаментов на биохимический путь метаболизма и физиологии [1]. Общий подход к созданию динамических моделей, описывающих процессы внутри и снаружи клеток, как правило, основан на построении системы нелинейных дифференциальных уравнений. Однако в настоящее время отсутствуют эффективные методы идентификации биохимических моделей, содержащих десятки параметров, по экспериментальным данным. Классические методы идентификации здесь практически неприменимы. Эволюционные алгоритмы работают очень долго. В связи с этим актуальна разработка новых методов и алгоритмов идентификации таких систем.

Анализ литературы. Идентификация параметров нелинейных биологических моделей является более сложной задачей, чем оценка параметров линейных моделей, потому что здесь не существует общего аналитического решения [1]. Биологические модели, как правило, являются динамическими и нелинейными, поэтому необходимо

использовать нелинейные оптимизационные подходы, где оценки рассогласование между прогнозами модели и экспериментальными данными используется в качестве критерия оптимальности, который необходимо минимизировать [2]. В связи с тем, что модели системной динамики имеют нелинейную природу и целевая функция, как правило, мультимодальна и не выпукла, градиентные методы оптимизации не могут быть применены для поиска глобального решения. В последние годы значительно возросла важность использования глобальных методов оптимизации для оценки параметров в системной биологии [3]. Глобальные методы оптимизации делятся на детерминистические, стохастические и гибридные. Некоторые детерминистические методы могут гарантировать, учитывая особенности конкретных задач, успешный поиск глобального оптимума. Отметим, что не существует детерминированных методов, которые могут решить проблему глобальной оптимизации, которая рассматривается в этой работе, за допустимое время. Их вычислительные затраты значительно увеличиваются (часто экспоненциально) при увеличении размерности задачи оптимизации. Этот класс алгоритмов не может быть применен для решения задач с большим количеством параметров. Алгоритмы, которые относятся к классу стохастических методов, основаны на вероятностных подходах. Их сходимость описана исключительно обобщенно на базе статистических данных. Однако, многие из стохастических методов могут найти решение близкое к глобальному оптимуму при приемлемом объеме вычислений. Стохастические методы не требуют также трансформации исходной задачи и могут рассматривать задачу оптимизации как черный ящик.

В данной работе проведено детальное тестирование работы генетических алгоритмов при оптимизации различных по сложности и количеству параметров искусственных целевых функций, и предложен механизм адаптации начальных параметров алгоритма в зависимости от сложности задачи [4]. В результате разработан универсальный алгоритм, который способен решать задачи любой сложности. Дополнительно предложен комбинированный метод на базе полученного ГА и локального детерминированного метода. Этот комбинированный метод позволяет улучшить точность решения и уменьшить вычислительные затраты. Полученный метод был интегрирован в моделирующую среду DIANA [5], которая разработана в институте Макса-Планка, г. Магдебург (Германия). Известны работы с применением параллельных ГА для решения задачи идентификации [6 – 9].

Целью работы является разработка и исследование предложенного гибридного метода для оптимизации сложной нелинейной биологической модели трехступенчатого биохимического метаболизма. Эта модель является сложной за счет большого количества параметров оптимизации (36), диапазон значений которых очень широк, и достаточно большого количества экспериментов (16), на базе которых выполняется идентификация.

Постановка задачи. В этой работе рассматривается задача идентификации параметров динамической модели при условии, что ее структура уже определена. Идентификация используется для поиска параметров модели, которые наиболее соответствуют набору экспериментальных данных. Для решения этой задачи разработан гибридный генетический алгоритм идентификации сложных нелинейных систем, сочетающий преимущества эволюционных и классических методов идентификации. Задача рассматривается на примере модели трехступенчатого биохимического метаболизма (three-step biochemical pathway), рис. 1. Эта задача была предложена Моулсом и его коллегами [7] как сложная тестовая задача идентификации параметров модели биохимического метаболизма с тремя энзиматическими шагами, которые непосредственно включают энзимы и mRNA. Эта проблема оптимизации рассматривалась в трудах Родригез Фернандеза и его коллег [3].

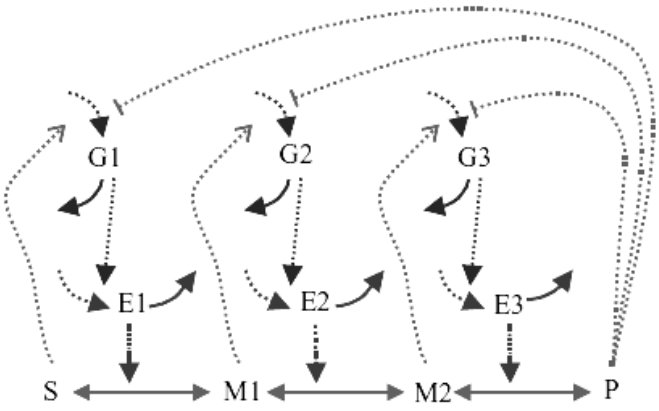

Рис. 1. Схема трехступенчатого биохимического метаболизма

Концентрации субстрата (S) и продукта (P) метаболизма остаються неизменными в течение всей реакции. M1 и M2 являются

промежуточными метаболитами; E1, E2 и E3 является энзимами; G1, G2 и G3 являются видами mRNA для энзимов. Непрерывные линии показывают реакции трансфера массы и указывают положительное направление потока, но они являються химически обратимыми. Пунктирные стрелки показывают активации, а пунктирные кривые с тупым концом показывают ограничение [7].

Проблема идентификации заключается в определении 36 кинетических параметров нелинейной биохимической динамической модели (описывается системой из 8 обыкновенных дифференциальных уравнений), которая описывает изменение концентраций метаболитов во времени [4, 6]).

Собственно идентификация сводится к проблеме оптимизации 36 параметров. Все параметры разделены на два разных класса: пиковые коэффициенты, которые изменяются в диапазоне (10^{-02}, 10^{+02}), и все остальные, которые изменяются в диапазоне (10^{-12}, 10^{+06}). Проблема глобальной оптимизации заключается в минимизации суммарного рассогласования J между экспериментальными и найденными значениями восьми переменных состояния, которые представлены вектором y

$$J = \sum_{i=1}^{n}\sum_{j=1}^{m}\{[(y_{pred}(i) - y_{exp}(i))_j]^2\},$$

где n – число данных для каждого эксперимента; m – число экспериментов; y_{exp} – экспериментальные данные; y_{pred} – вектор состояний, который был получен в результате моделирования с заданными значениями 36 параметров. Для лучшей оценки эффективности разработанного алгоритма были сгенерированы наборы псевдоэкспериментальних данных (в результате моделирования с номинальным набором параметров, поскольку этот оптимальный набор является известным). Итак, псевдозамеры концентраций метаболитов, протеинов и mRNA является результатом 16 различных экспериментов (моделирований), в которых значение субстрата S и продукта P были различными.

Генетический алгоритм. Для идентификации параметров указанной выше модели разработан гибридный генетический алгоритм оптимизации (ГА) на основе разработанного нами ранее ГА [4, 8]. Для этого алгоритма было проведено интенсивное исследование производительности при различных возможных параметрах ГА [4]. На базе этих исследований имплементирована автоматическая настройка параметров алгоритма в зависимости от сложности задачи. Параметры

выбираются таким образом, чтобы обеспечивать приемлемую точность решения при небольших (по сравнению с локальными и другими стохастическими методами) временных затратах на оптимизацию. Для повышения точности решения был разработан гибридный метод оптимизации, который сначала запускает ГА, а затем использует локальный детерминистический метод для улучшения результата [4]. Критерием остановки ГА было избрано формирование заданного процента схемы решения [9]. В большинстве случаев это не только приводит к повышению точности решения, но также уменьшает количество вычислений целевой функции. Проведена аналитическая и экспериментальная оценка эффективности гибридного метода на многих искусственных целевых функциях и на сложных динамических моделях, где он показал высокую эффективность [4].

Особенностью оптимизации таких моделей является их очень высокая сложность (36 параметров), мультимодальность, большое количество экспериментов (16), для которых выполняется идентификация параметров. Существенным также является то, что как указанной модели, так и многим другим биологическим моделям присуща высокая зависимость (чувствительность) параметров модели друг от друга [10]. В связи с этими особенностями были выполнены дополнительные модификации ГА. Используется логарифмическое кодирование (вместо обычного бинарного или кода Грея, которые применялись ранее [4]). Этот вид кодирования распределяет возможные значения параметра по логарифмической шкале, и точность также задается в логарифмической шкале. При этом несколько уменьшается абсолютная точность ГА (через меньшее количество возможных дискретных значений), однако увеличивается скорость работы ГА (требуется меньшее количество вычислений целевой функции для поиска решения). Для достижения необходимой точности используется гибридный метод оптимизации, который после окончания основного цикла ГА запускает локальный метод.

Локальный метод DN2GB[11] был использован вместо предыдущего градиентного метода. Предыдущий метод не может справиться с задачей такой сложности, особенно с учетом чувствительности параметров модели. Поэтому был выбран один из лучших локальных методов, использование которого является очень распространенным, к тому же он успешно используется для идентификации параметров сложных биологических моделей [10, 2]. Метод DN2GB был разработан Деннисом, Гайема и Уэлшем [11]. Этот алгоритм входит в библиотеку PORT [12], которая хорошо протестирована и эффективна, а также используется на практике для подобных типов задач. Этот алгоритм является вариацией

метода Ньютона и способен обрабатывать пределы параметров. Для эффективной работы ГА необходимо иметь полноценную начальную популяцию, поэтому был разработан специальный механизм формирования начальной популяции, при котором все элементы гарантированно будут находиться в области решений. Таким образом, ГА генерирует случайные значения для каждого элемента популяции, пока он не попадает в область решений. Элементы вне области решений, которые возникают (с достаточно высокой вероятностью) в процессе работы ГА, исключаются из популяции, при этом необходимый размер популяции будет восстановлен на этапе отбора. При решения задач оптимизации такой сложности присутствует достаточно высокая вероятность потери отдельных наиболее приспособленных элементов в результате отбора. В связи с этим в ГА был реализован механизм гарантированного перехода лучшего элемента в популяции в следующее поколение – элитизм [9].

Результаты. Разработанный гибридный алгоритм применен для оптимизации сложной биологической модели. Результаты исследований показали, что эффективность работы гибридного генетического алгоритма (GA + DN2GB) при решении той же задачи [13] выше эффективности как эволюционных стратегий со стохастическим подбором (SRES), так и гибридного метода на базе эволюционных стратегий и детерминированного метода DN2GB [10] (табл.).

Таблица

Сравнительные результаты применения методов оптимизации

Алгоритм	GA + DN2GB	SRES + DN2GB	SRES
Время работы	0.7	2.7	39.42
Количество вычислений	7.03E+4	1.89E+5	2.8E+6
Значение ЦФ	9.09E-09	1E-07	1.3E-03

За счет применения автоматической настройки параметров при инициализации ГА и модификаций разработанный алгоритм обеспечил точность решения задачи, сравнимую с точностью известных методов, которые используются для решения задачи оптимизации биологических моделей [13, 10]. При этом удалось сократить объем необходимых вычислений ЦФ, что уменьшает время работы алгоритма. Необходимо отметить, что алгоритм обеспечил решение задачи высокой сложности, и

при этом остался универсальным методом оптимизации. На рис. 2 представлен график изменения ЦФ в процессе работы гибридного ГА.

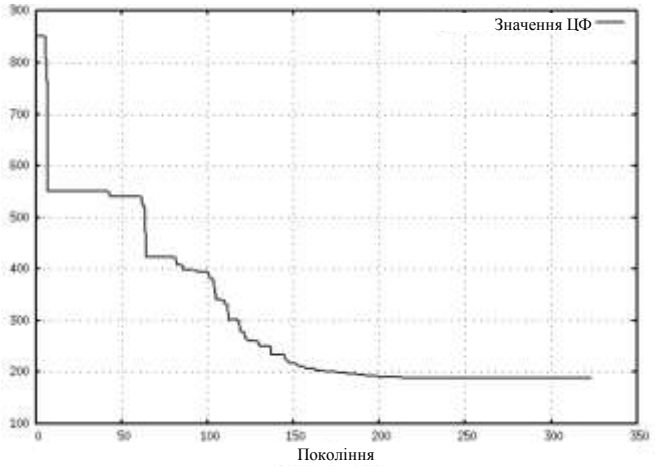

Рис. 2. Изменение значения ЦФ во время работы ГА в составе гибридного метода

Выводы. Проведено исследование разработанного метода для оптимизации сложной нелинейной биологической модели трехступенчатого биохимического пути метаболизма.

Результаты исследований показали, что эффективность работы разработанного гибридного генетического алгоритма выше, чем эффективность как эволюционных стратегий, которые используют стохастический отбор, так и гибридного метода на базе эволюционных стратегий и детерминированного метода. Необходимо отметить, что алгоритм обеспечил решение задачи высокой сложности и при этом остался универсальным методом оптимизации, который способен решать как простые задачи, так и оптимизировать мультимодальные нелинейные модели. Разработанный алгоритм интегрирован в моделирующую среду DIANA и используется для решения задач идентификации параметров сложных динамических систем, а именно химических и биологических моделей.

Список литературы: 1. *Kell D.B.* Metabolomics, modelling and machine learning in systems biology – towards an understanding of the languages of cells / *D.B. Kell* // FEBS Journal. – Prague. – 2006. – 273 (5). – P. 873-894. **2.** *Gennemark P.* Efficient algorithms for ordinary differential equation model identification of biological systems / *P. Gennemark, D. Wedelin* // Systems Biology, IET. – 2007. – Vol. 1. – Issue 2. – P. 120-129. **3.** *Rodriguez-Fernandez M.*

A hybrid approach for efficient and robust parameter estimation in biochemical pathways / *M. Rodriguez-Fernandez, P. Mendes, J. Banga* // 5th International Conference on Systems Biology. – 2006. – Vol. 83. – Issues 2-3. – P. 248-265. **4.** *Teplinskiy K.* Optimization Problems in the Technological-Oriented Parallel Simulation Environment / *K. Teplinskiy, V. Trubarov, V. Svjatnyj* // 18-th ASIM-Symposium Simulationtechnique SCS: Publishing House, Erlangen. – 2005. – P. 582-587. **5.** *Krasnyk M.* The ProMoT / Diana Simulation Environmen / *M. Krasnyk, K. Bondareva, O. Milokhov, K. Teplinskiy, M. Ginkel, A. Kienle* // 16th European Symposium on Computer Aided Process Engineering and 9th International Symposium on Process Systems. – Elsevier, Amsterdam, 2006. – P. 445-450. **6.** *Трубаров В.А.* Применение параллельного генетического алгоритма для решения задач оптимизации сложных динамических систем / *В.А Трубаров, К.С. Теплинський, И.В. Бабенко* // Научные труды Донецкого национального технического университета. – Серия "Проблемы моделирования и автоматизации проектирования динамических систем" (МАП-2007). – Донецк: ДонНТУ. – 2007. – Вип.: 6 (127). – С. 89-102. **7.** *Moles C.* Parameter estimation in biochemical pathways: a comparison of global optimization methods / *C. Moles, P. Mendes, J. Banga* // Genome Research. – 2003. – Vol. 13. – P. 2467-2474. **8.** *Трубаров В.А.* Подсистема оптимизации на базе эволюционных вычислений для параллельной моделирующей среды / *В.А.Трубаров, С.Ю. Гоголенко, К.С. Теплинский* // Региональная студеческая научно-техническая конференция "Компьютерный мониторинг и информационные технологии". – Донецк: ДонНТУ. – 2005. **9.** *Скобцов Ю.А.* Эволюционные вычисления: учебное пособие // *Ю.А. Скобцов, Д.В. Сперанский*. – М.: Национальный открытый университет "ИНТУИТ", 2015. – 331 с. 1**0.** *Rodriguez-Fernandez M.* Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems / *M. Rodriguez-Fernandez, J.A. Egea, J.R. Banga* // BMC Bioinformatics. – 2006. – Vol. 7. – P. 483. **11.** *Dennis J.* Algorithm 573, NLZSOL – An adaptive nonlinear least-squares algorithm / *J. Dennis, D. Gay, R. Welsch* // ACM Trans Math Software. – 1993. – Vol. 7. – P. 369-383. **12.** PORT Mathematical Subroutine Library [Electronic resource]. – URL: http://www.bell-labs.com/project/PORT/ **13.** *Runarsson T.* Stochastic ranking for constrained evolutionary optimization / *T. Runarsson, X. Yao* // IEEE Transactions on Evolutionary Computation. – 2000. **–** P. 284-294. **14.** *Фельдман Л.П.* Численные методы в информатике / *Л.П. Фельдман, А.І. Петренко, О.А. Дмитриева*. – К.: ВНР, 2004. – 420 с.

Bibliography (transliterated). 1. *Kell D.B.* Metabolomics, modelling and machine learning in systems biology – towards an understanding of the languages of cells / *D.B. Kell* // FEBS Journal. – Prague. – 2006. – 273 (5). – P. 873-894. **2.** *Gennemark P.* Efficient algorithms for ordinary differential equation model identification of biological systems / *P. Gennemark, D. Wedelin* // Systems Biology, IET. – 2007. – Vol. 1. – Issue 2. – P. 120-129. **3.** *Rodriguez-Fernandez M.* A hybrid approach for efficient and robust parameter estimation in biochemical pathways / *M. Rodriguez-Fernandez, P. Mendes, J. Banga* // 5th International Conference on Systems Biology. – 2006. – Vol. 83. – Issues 2-3. – P. 248-265. **4.** *Teplinskiy K.* Optimization Problems in the Technological-Oriented Parallel Simulation Environment / *K. Teplinskiy, V. Trubarov, V. Svjatnyj* // 18 th ASIM Symposium Simulationtechnique SCS: Publishing House, Erlangen. – 2005. – P. 582-587. **5.** *Krasnyk M.* The ProMoT / Diana Simulation Environmen / *M. Krasnyk, K. Bondareva, O. Milokhov, K. Teplinskiy, M. Ginkel, A. Kienle* // 16th European Symposium on Computer Aided Process Engineering and 9th International Symposium on Process Systems. – Elsevier, Amsterdam, 2006. – P. 445-450. **6.** *Trubarov V.A.* Primenenie parallel'nogo geneticheskogo algoritma dlja reshenija zadach optimizacii slozhnyh dinamicheskih sistem / *V.A. Trubarov, K.S. Teplins'kij, I.V. Babenko* // Nauchnye trudy Doneckogo nacional'nogo tehnicheskogo universiteta. – Serija "Problemy modelirovanija i avtomatizacii proektirovanija dinamicheskih sistem" (MAP-2007). – Doneck: DonNTU. – 2007. – Vip.: 6 (127). – P. 89-102. **7.** *Moles C.* Parameter estimation in biochemical pathways: a comparison of global optimization methods / *C. Moles, P. Mendes, J. Banga* // Genome Research. – 2003. – Vol. 13. – P. 2467-2474.

8. *Trubarov V.A.* Podsistema optimizacii na baze evoljuchionnyh vychislenij dlja parallel'noj modelirujushhej sredy / *V.A.Trubarov, S.Ju. Gogolenko, K.S. Teplinskij* // Regional'naja studecheskaja nauchno-tehnicheskaja konferencija "Komp'juternyj monitoring i informacionnye tehnologii". – Doneck: DonNTU. – 2005. **9.** *Skobcov Ju.A.* Jevoljucionnye vychislenija: uchebnoe posobie // *Ju.A. Skobcov, D.V. Speranskij*. – M.: Nacional'nyj otkrytyj universitet "INTUIT", 2015. – 331 s. **10.** *Rodriguez-Fernandez M.* Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems / *M. Rodriguez-Fernandez, J.A. Egea, J.R. Banga* // BMC Bioinformatics. – 2006. – Vol. 7. – P. 483. **11.** *Dennis J.* Algorithm 573, NLZSOL – An adaptive nonlinear least-squares algorithm / *J. Dennis, D. Gay, R. Welsch* // ACM Trans Math Software. – 1993. – Vol. 7. – P. 369-383. **12.** PORT Mathematical Subroutine Library [Electronic resource]. – URL: http://www.bell-labs.com/project/PORT/ **13.** *Runarsson T.* Stochastic ranking for constrained evolutionary optimization / *T. Runarsson, X. Yao* // IEEE Transactions on Evolutionary Computation. – 2000. – P. 284-294. **14.** *Fel'dman L.P.* Chislennye metody v informatike / *L.P. Fel'dman, A.I. Petrenko, O.A. Dmitrieva*. – K.: VNR, 2004. – 420 p.

Поступила (received) 10.08.2015

Статью представил д.т.н., проф. ДонНТУ Скобцов Ю.А.

Teplinskiy Konstantin, master
postgraduate student
Donetsk National Technical University
Sq. Shybankova, Krasnoarmiysk, Donetsk region, Ukraine, 85300
tel./phone: +49 (172) 4599424, e-mail: konstantin.teplinskiy@gmail.com
ORCID ID: 0000-0003-4204-7547