Компьютерная математика

В.О. РУДИК

ПРО ПІДХОДИ ДО РОЗРОБКИ АЛГОРИТМІВ ОМК ДЛЯ РОЗВ'ЯЗУВАННЯ ЗАДАЧІ ПЕРЕДБАЧЕННЯ СТРУКТУРИ ПРОТЕЇНІВ

Проводиться огляд найбільш відомих алгоритмів методу оптимізації мурашиними колоніями (ОМК), які розроблені для пошуку оптимальної структури протеїну для HP моделі Ділла. Описано відмінні риси цих алгоритмів на різних етапах схеми, що використовуються авторами робіт для врахування особливостей задачі. Наведено різні підходи до розпаралелювання процедури з використанням центральних та графічних процесорів. Наведено приклади гібридних алгоритмів, складовою частиною яких є метод ОМК.

УДК 519.8

Вступ. Протеїни – це важливі частини нашого життя, вони беруть участь практично у кожному життєвому процесі в клітинах. Розуміння білкових структур є життєво важливим для визначення функцій білка, тому дослідження процесу згортання протеїну – одна з найбільш важливих задач у галузі біоінформатики і теоретичної хімії. Результати цих досліджень вносять вагомий вклад у медицину (наприклад, для дизайну ліків) та біотехнології (наприклад, при розробці нових ферментів). Класичні методи – рентгенівська кристалографія та ядерний магнітний резонанс є дорогими і можуть зайняти багато часу (іноді більше року). Тому широкого використання у цій сфері набули обчислювальні методи.

Однією з найбільш широко використовуваних моделей процесу згортання є гідрофобно-полярна (HP) модель Ділла [1]. Проблема передбачення нативної структури у ній зводиться до спеціальної задачі комбінаторної оптимізації. Перспективні підходи в її розв'язуванні базуються на методах оптимізації мурашиних колоній (ОМК) [2]. В роботі проводиться огляд ряду поширених алгоритмів, що розроблені на базі цього методу.

Гідрофобно-полярна модель Ділла. Молекула протеїну представляє собою послідовність амінокислотних залишків, поєднаних пептидними зв'язками у ланцюг. Задача передбачення третинної структури протеїну

полягає у побудові просторової форми молекули виходячи з відомої послідовності залишків. У HP моделі Ділла ця форма подається шляхом без самоперетинів у обраних дискретних ґратках, так що сусідні в послідовності залишки розташовуються в сусідніх вузлах ґраток. Амінокислотні залишки в залежності від їх властивостей поділяються на гідрофобні і полярні, і енергія структури рахується як кількість гідрофобних залишків, розташованих у сусідніх вузлах ґраток, але не сусідніх у послідовності (HH-контакти), зі знаком мінус. Вважається, що молекула приймає ту форму, яка мінімізує її енергію, таким чином формулюється задача комбінаторної оптимізації. Різноманітні алгоритми запропоновані для її розв'язання, серед яких чільне місце займають алгоритми методу ОМК.

ОМК. Ідея методу ОМК запозичена при спостереженні за колонією мурах, які шукають найкоротший шлях від мурашника до їжі. Кожна мураха залишає на своєму шляху феромонний слід і при побудові маршруту враховує феромонні сліди, відкладені на ньому. Таким чином з часом найкоротші маршрути помічаються більшою кількість феромону і мурахи обирають їх.

Опишемо цей процес більш формально. Розглянемо задачу комбінаторної оптимізації [3], простором розв'язку якої є впорядковані послідовності довжини n із елементів множини $A = \{a_1, ..., a_m\}$ (позначимо його A^n), а цільовою функцією, яку потрібно мінімізувати – функція $E : A^n \to \mathbb{R}$, де $\mathbb{R}$ позначає множину дійсних чисел. Загальна обчислювальна схема алгоритму ОМК показана на рисунку.

procedure ACO()
  $r_{rec} := null$
  Ініціалізувати феромонну матрицю (Φ)
  while not *Умова завершення* **do**
    for $i = 1,\ldots, N$ **do**
      $r_i :=$ *Згенерувати розв'язок* (Φ)
    end for;
    Застосувати локальний пошук $(r_1, ..., r_N)$
    for $i = 1,\ldots, N$ **do**
      Поновити феромонну матрицю (Φ, r_i)
    $r_{rec} := \arg \min_{r \in \{r_{rec}, r_1, ..., r_N\}} \{E(r)\}$
  end while
  return r_{rec}
end procedure.

РИСУНОК. Загальна схема алгоритму ОМК

Тут N – кількість мурах у популяції. Метод є ітеративним, і інформація про простір накопичується і передається з ітерації в ітерацію через матрицю Φ розмірністю $m \times n$, що зветься феромонною матрицею. На кожній ітерації алгоритму генерується популяція допустимих розв'язків, кожен з яких формується на основі використання випадкового розподілу (функція згенерувати розв'язок (Φ) на схемі):

$$p(r_i = a_j) = \frac{\tau_{ij}^{\alpha} \eta_{ij}^{\beta}}{\sum_{k=1}^{m} \tau_{ik}^{\alpha} \eta_{ik}^{\beta}}, \quad (1)$$

де $r = (r_1, ..., r_n) \in A^n$ позначає розв'язок, $p(\cdot)$ – функцію ймовірності, τ_{ij} – значення елемента феромонної матриці у i рядку, j стовпчику (його називають інтенсивністю феромонного сліду), а η_{ij} – евристичні значення, які задаються з урахуванням специфіки задачі і відображають «бажаність» відповідного елемента в розв'язку. Параметри алгоритму α та β задають відносну важливість феромонної та евристичної інформації і обираються з емпіричних міркувань або базуючись на аналізі обчислювального експерименту.

Після генерації популяції до всіх або частини її елементів зазвичай застосовується локальний пошук, хоча цей крок не є обов'язковим. Далі проводиться поновлення значень елементів феромонної матриці за схемою:

$$\tau_{ij} := (1 - \rho)\tau_{ij} + \sum_{r: r_i = a_j} \Delta(r),$$

де $\Delta(r)$ відображає відносну якість розв'язку r у популяції та забезпечує підкріплення феромонного сліду, а параметр ρ, що називається ступенем випаровування феромону і лежить у інтервалі $(0,1)$, визначає, наскільки швидко інформація, зібрана на попередніх ітераціях, «забувається».

Ітеративний процес продовжується, поки не виконується умова зупинки. Найоптимальніший (за значенням цільової функції) з розв'язків, згенерованих у процесі роботи алгоритму, повертається як результат.

Доведено, що при виконанні певних умов реалізації алгоритми методу ОМК збігаються з ймовірністю 1 [4].

Алгоритми методу ОМК для розв'язування задачі прогнозування третинної структури протеїнів. Для розробки ефективних алгоритмів методу ОМК необхідно враховувати специфіку розв'язуваної задачі. Розглянемо, які схеми запропоновано в літературі для розв'язання цим методом задачі прогнозування третинної структури молекули протеїну у HP моделі Ділла.

В першу чергу необхідно задати простір розв'язків, які будуть кодувати третинну структуру протеїну у формі, що використовується у схемі ОМК, тобто визначити множину $A = \{a_1, ..., a_m\}$. Нагадаємо, що розв'язком задачі прогнозування третинної структури протеїну є шлях без самоперетинів у деяких дискрет-

них ґратках, тобто послідовність їх координат, яка задовольняє умові зв'язності та відсутності самоперетинів. Зрозуміло, що форма множини A буде залежати від типу обраних ґраток. В літературі зустрічаються, зокрема, двовимірні квадратні та трикутні [5 – 11], [9], тривимірні кубічні [5], [9], [12 – 16] та трикутні [9], [17] ґратки. Однак на практиці використовувати як множину A координати ґраток є неефективним, оскільки це призводить до надмірних розмірів феромонної матриці. В алгоритмах ОМК зручно задавати шлях через його абсолютне кодування – зсув наступного залишку відносно попереднього [14], [16]. Для двовимірних ґраток також використовується подання шляху через відносне кодування – через кути між позиціями амінокислотних залишків. У роботі [17] запропоновано варіант побудови відносного кодування для тривимірних трикутних ґраток. Приклад іншого кодування наведено у [15]. Тут будуються 150 фрагментів, що покривають усі допустимі структури довжини 4 у кубічних ґратках і шлях кодується послідовністю таких структур.

Принципово інший підхід використовується у [6]. Тут феромон відкладається прямо на ребра ґраток, що задані в постановці задачі.

Для ініціалізації феромонної матриці переважно використовуються стандартні для ОМК схеми: ініціалізація однаковими або випадковими значеннями. Кількість елементів у популяції варіюється від 5 до 1500, і навіть 8192 при використанні паралельних алгоритмів на графічних процесорах [18]. Зокрема, у [5] розглядалися популяції розміром від 1 до 5000. Найефективнішими виявились колонії із 100 агентів як для двовимірного, так і для тривимірного випадку. Також у [5] показано, що запропонований алгоритм стійкий до змін розміру популяції, тільки у випадках занадто малих або занадто великих колоній його ефективність знижується.

Побудова розв'язку починається з першого елемента [13, 14], з центрального елемента [6] або з випадково вибраного елемента послідовності рівноймовірно для всіх елементів [5, 7, 8, 12, 16]. Якщо побудова починається з випадкового елемента, шлях може розвиватись у обох напрямках. Використовуються різні варіанти вибору напрямку:

а) напрямок вибирається так, щоб пропорція недобудованих амінокислотних залишків справа і зліва залишалась (приблизно) незмінною;

б) обирається напрямок, у якому більше недобудованих залишків;

в) напрямок вибирається випадково з імовірністю, пропорційною кількості недобудованих залишків у цьому напрямку.

Найбільш широко використовуваною процедурою для визначення евристичних значень η_{ij} є задання їх як

$$\eta_{ij} = 1 + h_{ij}, \tag{2}$$

де h_{ij} – кількість HH контактів, що будуть створені при позиціонуванні амінокислоти i у напрямку j. У роботі [9] ця формула модифікується для адаптації під ширше коло ґраток:

$$\eta_{ij} = 1 + h_{ij} + \frac{1}{2} h_{ij}^{\,2}.$$

У роботах [14, 16] використовується екстремальний варіант:

$$\eta_{ij} = h_{ij},$$

при такому визначенні для всіх полярних залишків $\eta_{ij} = 0$ і ймовірність $p(r_i = a_j)$ приймається однаковою для всіх допустимих j, тобто феромонна матриця впливає виключно на позиціонування гідрофобних залишків.

У роботі [6] евристичні значення для гідрофобних залишків рахуються за формулою (2), а для полярних – за формулою

$$\eta_{ij} = v_{ij} + h_{ij}' + 1,$$

де v_{ij} та h_{ij}' – кількості вільних вузлів і полярних амінокислотних залишків у околі вузла, який розглядається для позиціонування залишку на поточному кроці. У роботі [17] пропонується узагальнення цього варіанта шляхом введення додаткових параметрів.

Інший варіант – підрахунок значень η_{ij} у відповідності з розподілом Больцмана [5, 7, 10]:

$$\eta_{ij} = e^{-\gamma h_{ij}},$$

де γ – параметр, що називається зворотною температурою.

У роботі [7] при побудові структури додатково вводяться певні емпіричні обмеження, мета яких – заборонити неефективні з емпіричної точки зору вибори мурах.

Специфічною для задачі прогнозування третинної структури протеїнів є проблема потрапляння в глухий кут на етапі побудови розв’язку – ситуації, коли наступний елемент послідовності не може бути доданий без порушення умови відсутності самоперетинів. Ця проблема є особливо актуальною у двовимірних ґратках. Для її вирішення використовується процедура відкату на один крок [7, 8, 17], на половину побудованої згортки [5] чи на випадкову кількість кроків [6]. Після відкату останній напрямок, який було вилучено, помічається як табу і решта шляху добудовується відповідно до процедури побудови розв’язку (1) або ж випадковим чином. У роботі [13] для виходу з глухих кутів запропонована схема, подібна тій, що використовується для розв’язання колізій, що виникають коли два комп’ютери одночасно надсилають інформацію в мережі Ethernet. Схема наступна: спочатку робиться відкат на 2 кроки, якщо все одно виникає глухий кут, відкатуються 4 кроки, якщо і це не допомогло – 8, у випадку невдачі – знову 8. Нарешті, якщо з глухого кута вийти не вдалось, шлях починають будувати з самого початку.

У роботі [10] відкат не використовується, замість цього структура, що потрапила в глухий кут, заміняється частиною структури, яка є найоптимальнішою у популяції на поточному кроці. Мурахи при такій схемі будують розв’язки синхронно, а не послідовно.

Враховуючи трудомісткість процедур локального пошуку для розв'язуваної задачі, використовувати їх для всіх елементів популяції недоречно. Локальним пошуком покращують частку (5 % – 50 %) найоптимальніших елементів популяції [5, 8]. Локальний пошук проводиться з повтореннями – його результати додаються у популяцію без заміни структур, які вони покращили, при цьому використовується 1 % найкращих елементів популяції. У роботах [6, 14] локальний пошук не використовується.

Серед запропонованих схем локального пошуку поширені такі.

Локальний пошук з мутацією далекобійним переміщенням [5]. Схема мутації далекобійним переміщенням така: відповідно до рівномірного розподілу вибирається елемент структури і його напрямок змінюється на випадковий (рівноймовірно для всіх). Після цього кожен наступний елемент з ймовірністю $\hat{p}$ не змінює свій напрямок, якщо той є допустимим. У протилежному випадку його напрямок встановлюється випадково серед допустимих з ймовірністю, пропорційною евристичному значенню η_{ij}. Локальний пошук проводиться мутаціями далекобійним переміщенням до першого покращення, якщо за певну кількість мутацій покращення не було знайдено, процедура зупиняється. Переваги такої схеми в тому, що вона майже не призводить до недопустимих розв'язків навіть у випадку, коли згортка є досить компактною, вона розглядає більший окіл, ніж звичайна одноточкова мутація, а також вона є пояснювальною з фізичної точки зору. У роботі [12] як локальний пошук використовується окремий випадок описаної схеми ($\hat{p} = 1$).

Локальний пошук з точковою та парною мутацією [8]. Для точкової чи парної мутації випадковим чином обирається індекс амінокислотного залишку l , після чого напрямки у шляху в позиції l при точковій і в позиціях l та $l+1$ при парній мутації змінюються на такі, які покращують енергію структури, якщо вони існують.

Локальний пошук з заміною шаблонів: у роботі [13] сусіди для локального пошуку формуються зміною визначеної шаблонної частини згортки на іншу задану. Наводиться три приклади таких перетворень, які названі «зсув по діагоналі», «колінчастий поворот» та «рух змії». До цієї ж категорії відноситься «рух затяжкою», що використовується в [7, 10].

При поновленні феромонної матриці значення $\Delta(r)$ задаються як відносна якість розв'язку r [5, 9]:

$$\Delta(r) = \frac{E(r)}{E(r^*)},$$

де r^* – оптимальний серед знайдених на поточний момент варіанта розв'язку. У роботі [17] запропонована процедура поновлення по частинах структур, які найбільше впливають на їх енергію.

У деяких алгоритмах використовується елітарний підхід – поновлення проводять тільки по частині найоптимальніших розв'язків, або навіть по одному

найкращому із згенерованих на поточній ітерації. У роботі [16] розглядаються також альтернативні функції енергії, які можуть бути використані в алгоритмі для того, щоб спрямувати евристику більш ефективним шляхом у процесі пошуку розв'язків задачі.

Окрім глобального поновлення феромону, яке проходить після того, як популяція побудована, представлені також схеми з локальним поновленням – в процесі генерації популяції. Так у [13] пропонується підсилювати феромонні сліди, що відповідають структурам з найнижчою енергією серед згенерованих на поточній ітерації, щоб пошук у просторі проходив ближче до оптимальніших значень. У роботах [6, 14, 16] використовується протилежний підхід – з метою диверсифікації пошуку після того, як мураха будує маршрут, феромонний слід на ньому випаровується, щоб наступна мураха не пройшла по тому ж маршруту.

У частині робіт використовується підхід з MAX-MIN Ant System [4] для ренормалізації значень феромонної матриці [5, 13].

Умовою завершення роботи алгоритму є виконання фіксованої кількості ітерацій або фіксованої кількість ітерацій без покращень. У роботі [8] алгоритм зупиняється після того, як на протязі трьох ітерацій не відбувалось покращення оптимального значення, після чого в три рази зменшується значення параметра α у формулі (1) і робиться рестарт алгоритму.

Розробка паралельних алгоритмів ОМК. Схема методу ОМК легко розпаралелюється. Зокрема ряд паралельних алгоритмів запропоновано і для розв'язування розглядуваної задачі. Опишемо декілька схем розпаралелювання.

Схема Distributed Single Colony (DSC) [12]: феромонна матриця зберігається на центральному сервері, а клієнти після фаз побудови і локального пошуку надсилають на нього обрані структури; за цими структурами на центральному сервері відбувається поновлення значень феромонної матриці і її копія передається клієнтам для наступної ітерації. Аналогічна схема реалізована в [8].

Схема Distributed Multi Colony with Circular Exchange of Migrants (MCOCE) [12]: кожен підлеглий процес розглядається як незалежна колонія. Після кожних k ітерацій феромонна матриця колоній процесів поновлюється агентами сусідніх колоній, при цьому структура сусідства має форму кола.

Схема Distributed Multi Colony with Pheromone Matrix Sharing (MCOPE) [12]: кожен підлеглий процес реалізує незалежну колонію, через кожних k ітерацій всі феромонні матриці поновлюються за формулою:

$$\tau_{cij} := \tau_{cij} + \frac{\left(\sum_{l=1}^{C} \tau_{lij} - \tau_{cij}\right)\gamma}{C-1},$$

де C – кількість підлеглих процесорів, c – номер процесора, а γ – параметр алгоритму.

У роботі [18] розроблено алгоритм для використання на графічних процесорах (CUDA). Тут один потік асоціюється з одною мурахою. Відкладення феромонних слідів також проводиться паралельно, при чому на цьому кроці виділяється окремий потік для кожної амінокислоти, і при реалізації використовуються атомні інструкції.

Гібридні алгоритми. Для розв'язання задачі прогнозування третинної структури протеїнів ідеї методу ОМК застосовуються також в комбінації з ідеями інших методів при формуванні гібридних алгоритмів. Так, в [7] пропонується поєднання методу ОМК з методом гілок і меж, в [15] – з методом Монте Карло з марківськими ланцюгами, в [11] – з методом навчання з підкріпленням.

Висновки. Проведено огляд ряду поширених алгоритмів методу ОМК, розроблених для розв'язування задачі передбачення третинної структури протеїнів. Особливістю задачі є відносно високі затрати часу на пошук допустимого розв'язку і процедуру локального пошуку та проблема глухих кутів при побудові згорток. Авторами алгоритмів розроблено різні схеми для врахування цих специфік. Основними відмінностями між алгоритмами є спосіб подання шляху у ґратках, визначення евристичного значення при побудові шляху, варіанти виходу з глухих кутів, процедури локального пошуку. Кожен з поданих алгоритмів показав свою ефективність при розв'язуванні певного кола задач передбачення просторової структури протеїнів.

В.А. Рудык

О ПОДХОДАХ К РАЗРАБОТКЕ АЛГОРИТМОВ ОМК ДЛЯ РЕШЕНИЯ ЗАДАЧИ ПРОГНОЗИРОВАНИЯ СТРУКТУРЫ ПРОТЕИНОВ

Проводится обзор ряда наиболее известных алгоритмов метода оптимизации муравьиными колониями, что разработаны для поиска оптимальной структуры в HP модели Дилла. Описано отличные черты этих алгоритмов на различных этапах схемы, что используются авторами работ для учета особенностей задачи. Приведены различные подходы к распараллеливанию процедуры с использованием центральных и графических процессоров. Приведены примеры гибридных алгоритмов, частью которых является метод ОМК.

V.O. Rudyk

ABOUT THE APPROACHES TO ACO ALGORITHMS DEVELOPMENT FOR SOLVING PROTEIN STRUCTURE PREDICTION PROBLEM

An overview of a number of the most popular ant colony optimization algorithms that are developed for finding the optimal conformation for HP model is presented. Specific features of these algorithms at different stages of the scheme are described and used by the authors to take into account the peculiarities of the problem. Various approaches to parallelization of the procedure for multicore CPU and GPU support are given. Examples of hybrid algorithms with ACO as a part of them are provided.

1. Dill K., Bromberg S., Yue K., Fiebig K., Yee D., Thomas P., Chan H. Principles of protein folding – a perspective from simple exact models. *Protein Science*. 1995. N 4. P. 561 – 602.
2. Dorigo M., Stützle T. Ant colony optimization. Cambridge (MA): MIT Press, 2004. 348 p.
3. Гуляницький Л.Ф., Мулеса О.Ю. Прикладні методи комбінаторної оптимізації. К.: Видавничо-поліграфічний центр "Київський університет", 2016. 142 с.

4. Hoos H., Stutzle T. Stochastic Local Search: Foundations and applications. San Francisco: Morgan Kaufmann Publ., 2005. 658 p.
5. Shmygelska A, Hoos H. An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem. *BMC Bioinformatics.* 2005. **30**, N 6. P. 30 – 52.
6. Hu X., Zhang J., Xiao J., Li Y. Protein Folding in Hydrophobic-Polar Lattice Model: A Flexible Ant Colony Optimization Approach. *Protein and Peptide Letters*. 2008. **15**, N 5. P. 469 – 477.
7. Thalheim T., Merkle D., Middendorf M. Protein Folding in the HP-Model Solved with a Hybrid Population Based ACO Algorithm. *IAENG International Journal of Computer Science*. 2008. **35**, N 3. P. 291 – 300.
8. Guo H., Lu Q., Wu J., Huang X., Qian P. Solving 2D HP Protein Folding Problem by Parallel Ant Colonies. *Biomedical Engineering and Informatics.* 2nd International Conference. 2009. P. 1–5.
9. Nardelli M., Tedesco L., Bechini A. Cross-lattice Behavior of General ACO Folding for Proteins in the HP Model. *Proceedings of the 28th Annual ACM Symposium on Applied Computing*. 2013. P. 1320 – 1327.
10. Liu Z., Yang Z. Heuristic Ant Colony Optimization Algorithm for predicting The Structures of 2D HP Model Proteins. *Biomedical Engineering and Informatics*. 7th International Conference. 2014. P. 719 – 723.
11. Dogan B., Olmez T. A novel state space representation for the solution of 2D-HP protein folding problem using reinforcement learning methods. *Applied Soft Computing*. 2015. N 26. P. 213 – 223.
12. Chu D., Zomaya A. Parallel Ant Colony Optimization Algorithms for the 3D Protein Structure Prediction using the HP Lattice Model. *Parallel Evolutionary Computations*. 2006. P. 177 – 198.
13. Song J., Cheng J., Zheng T. Protein 3D HP model folding simulation based on ACO. *Intelligent Systems Design and Applications*. 2006. P. 410 – 415.
14. Fidanova S., Lirkov I. Ant Colony System Approach for Protein Folding. Int. Conf. *Multiconference on Computer Science and Information Technology*. 2008. P. 887 – 891.
15. Citrolo A., Mauri G. A Hybrid Monte Carlo Ant Colony Optimization Approach for Protein Structure Prediction in the HP Model. *Electronic Proceedings in Theoretical Computer Science*. 2013. N 130. P. 61 – 69.
16. Thilagavathi N., Amudha T. ACO-Metaheuristic for 3D-HP Protein Folding Optimization. *ARPN Journal of Engineering and Applied Sciences*. 2015. **11**, N 10. P. 4948 – 4953.
17. Рудык В. Анализ сходимости алгоритма метода оптимизации муравьиными колониями для решения задачи предсказания структуры протеина. *Журнал обчислювальної та прикладної математики*. 2013. № 4. С. 48 – 54.
18. Llanes A., Velez C., Sanchez A., Perez-Sanchez H., Cecilia J. Parallel Ant Colony Optimization for the HP Protein Folding Problem. *International Conference on Bioinformatics and Biomedical Engineering*. 2016. P. 615 – 626.

Одержано 31.10.2017

Про автора:

Рудик Віталіна Олександрівна,
молодший науковий співробітник
Інституту кібернетики імені В.М. Глушкова НАН України.
E-mail: vitalina.rudyk@gmail.com