ТЕОРІЯ ОПТИМАЛЬНИХ РІШЕНЬ

УДК 519.21

В.А. РУДЫК

АНАЛИЗ ЭФФЕКТИВНОСТИ АЛГОРИТМОВ ОМК ДЛЯ РЕШЕНИЯ ЗАДАЧИ О СВОРАЧИВАНИИ ПРОТЕИНОВ

Рассматривается гидрофобно-полярная модель сворачивания протеина, предлагается и исследуется алгоритм прогнозирования третичной структуры белка, построенный на базе метода ОМК. Строится модель молекулы протеинов в трехмерной треугольной решетке. Разработаны методы априорной оценки позиции в решетке и обновления матрицы феромонных путей, их эффективность подтверждена вычислительным экспериментом.

Введение. В современном мире в разных отраслях науки с открытием неизвестных ранее объектов возрастает необходимость разработки новых методов для их изучения. Важным становится обмен информации между различными науками для применения идей из одной сферы в другой. Так, например, в прикладной математике оказалось эффективным использование генетических алгоритмов и искусственных нейронных сетей, которые имитируют процесс эволюции и процессы, протекающие в мозге, соответственно. В то же время в биологии для понимания различных механизмов и процессов часто недостаточно исключительно экспериментальных данных. Это привело к появлению различных междисциплинарных исследований на стыке прикладной математики, информатики, биологии и химии, таких как биоинформатика, математическая биология, вычислительная биология, биокибернетика, структурная геномика, биоинженерия. В настоящее время эти области являются одними из наиболее перспективных и развивающихся.

Среди открытых задач вычислительной биологии – определение третичной структуры протеина [1,2]. Разнообразие белковых молекул делает невозможным решение проблемы путем проведения экспериментов, но построив на основе наблюдений математическую модель сворачивания молекулы можно спрогнозировать ее форму. Ниже используется одна из таких моделей – дискретная HP-модель Дилла [1].

1. Прогнозирование структуры протеина как задача комбинаторной оптимизации. Исходными данными для определения третичной структуры белка – формы, которую принимает молекула в трехмерном пространстве – является его первичная структура, которая представляет собой линейную последовательность аминокислот. Одна из важных характеристик аминокислотных остатков – гидрофобность – выражает, насколько он предпочитает неполярное окружение (например, этанол в качестве растворителя или внутренность белка) полярному (например, воде). В зависимости от этой характеристики все аминокислоты делятся на два класса – гидрофобные и полярные (или гидрофильные). Попадая в воду белок, являя собой цепочку аминокислот, принимает ту форму, в которой площадь поверхности контакта воды с полярными аминокислотами максимизируется, а с гидрофобными минимизируется. Поскольку в результате взаимодействия образуется шарообразная структура, такие белки называют глобулярными.

Дискретной моделью формы молекулы протеина является путь без самопересечений в некоторой дискретной решетке. В исследованиях чаще всего встречаются двумерная квадратная [2], треугольная [3] и трехмерная кубическая [2, 4] решетки, рассматриваются и другие виды решеток, а также дискретных нерешеточных структур [3,5]. Соседние в цепочке аминокислотные остатки располагаются в соседних узлах решетки, таким образом соблюдается условие неразрывности молекулы. Между гидрофобными остатками, которые располагаются в соседних узлах решетки, но не являются соседями в цепочке, образуются гидрофобные связи. В модели Дилла в качестве значения свободной энергии структуры принято считать количество таких связей в ней со знаком минус. Соответствуя законам термодинамики, молекула образует структуру с минимальной свободной энергией. Это дает возможность формализовать проблему как задачу комбинаторной оптимизации.

Для представления пути в дискретной решетке можно использовать один из трех методов: последовательность координат узлов аминокислотных остатков, абсолютное и относительное кодирование [6]. Мы будем рассматривать последние два представления, поскольку первое требует постоянных проверок неразрывности молекулы, а второе и третье обеспечивают связность автоматически.

2. Метод оптимизации муравьиными колониями (ОМК). Исследование дискретных моделей естественных процессов часто приводит к NP-сложным задачам; на практике это значит, что в общем случае найти решение за допустимое (полиномиальное) время невозможно (согласно гипотезе о том, что классы P и NP не равны). Одной из таких задач является сформулированная выше на основе HP-модели задача комбинаторной оптимизации. Парадокс заключается в том, что прообразы в природе таких моделей часто решают ту же задачу практически мгновенно – некоторые белки сворачиваются за миллионную долю секунды. Этот факт побуждает исследователей строить цепочку решения задачи по аналогии с некоторыми естесственными процессами. Кроме уже упомянутых генетических алгоритмов и нейронных сетей к алгоритмам, имеющим аналоги в

природе, относятся метод имитации отжига, иммунные алгоритмы, алгоритм ОМК. Рассмотрим последний подробнее.

В поисках пищи муравьи постоянно решают оптимизационную задачу – поиск кратчайшего пути от муравейника до еды. Для этого они используют феромоны, помечая ими свой путь. Выбирая направление движения, муравей ориентируется на феромонную дорожку, и чем она сильнее, тем вероятнее, что муравей последует по ней. В конечном итоге в большинстве случаев все рабочие муравьи движутся по одному субоптимальному пути от муравейника до еды и обратно. В контексте задачи комбинаторной оптимизации это сводится к популяционному алгоритму, на каждой итерации которого все элементы популяции – агенты – пошагово строят решения задачи. Информация об оптимальности значения целевой функции на этих решениях записывается в феромонную матрицу, которую учитывают агенты на следующей итерации.

Алгоритмы метода ОМК целесообразно применять для поиска приближенных решений NP-сложных задач, в том числе и задачи о сворачивании протеина. Так, в [2] с его помощью строится структура молекулы в квадратной и кубической решетке, предложена процедура локального поиска для повышения показателей эффективности алгоритма. В [4] описано решение, использующее двухшаговое обновление феромонных путей на каждой итерации, рассматривается трехмерная кубическая решетка.

Ниже предложен алгоритм для решения задачи в трехмерной кубической решетке. Разработан отличный от ранее описанных подход к обновлению феромонной матрицы и подсчета априорных оценок.

3. Описание алгоритма. На вход задачи подается первичная структура протеина, которая в терминах изложенной модели принимает вид $o_1 o_2 ... o_N$, $o_i \in \{H, P\}$, $i = \overline{1, N}$, где N обозначает количество аминокислот в молекуле, а H и P – гидрофобный и полярный остаток в цепочке соответственно. Для исследования предлагается использовать трехмерную треугольную решетку [6]. У каждого узла такой решетки 12 соседей. В разработанном алгоритме можно применять как абсолютную, так и относительную кодировку. Закодированное решение имеет вид последовательности векторов $r_1 r_2 ... r_{N-1}$, $r_i \in A$, $i = \overline{1, N-1}$, где A – множество значений элементов кода [6], зависящее от вида решетки и типа кодировки. Без ограничения общности можно рассматривать только кодировки $v = r_2 r_3 ... r_{N-1}$ (без первого элемента), поскольку r_1 определяет поворот структуры, а форма молекулы задается путем в решетке с точностью до поворота. Количество элементов множества A обозначим символом n, а сами элементы – $dir_1, dir_2, ..., dir_n$ (от слова «direction» – направление).

Обозначим феромонную матрицу символом Φ – в нашей формулировке задачи она будет иметь размер $n \times (N-2)$. Схема алгоритма ОМК, разработанного и исследованного для нашей задачи, приведена на рис.1. В нем параметр N_P задает размер популяции, а N_{LS} – количество элементов, для которых использует-

ся локальный поиск. Первый шаг – *ИнициализироватьФеромоннуюМатрицу*(Ф) – инициализация матрицы случайными значениями в диапазоне от 0 до 1.

```
procedure ACO()
    fold_rec := null;
    ИнициализироватьФеромоннуюМатрицу( Ф );
    while (not УсловиеЗавершения()) do
        for i = 1,…, N_P do
            fold_i :=ДопустимоеРешение( Ф );
        end for;
        Отсортировать( fold_i );
        for i = 1,…, N_LS do
            fold_i :=Локальный поиск( fold_i );
        end for;
        fold_rec :=ОптимальноеЗначение( fold_rec , fold_1 ,…, fold_N_LS );
        ИспаритьФеромоннуюМатрицу( Ф );
        for i = 1,…, N_P do
            ОбновитьФеромоннуюМатрицу( Ф , fold_i );
        end for
    end while;
    return fold_rec ;
end procedure.
```

Рис. 1. Схема алгоритма ОМК для задачи определения структуры протеина

Процедура *ДопустимоеРешение*(Ф), используя феромонную матрицу строит путь без самопересечений, пошагово добавляя к кодировке элементы с вероятностью:

$$P(r_i = dir_j) = \frac{\Phi[j,i\text{-}1]^{\alpha} * \mathrm{Est}(r_2 r_3 \ldots r_{i-1}, dir_j)^{\beta}}{\sum_{j=1}^{n} \Phi[j,i\text{-}1]^{\alpha} * \mathrm{Est}(r_2 r_3 \ldots r_{i-1}, dir_j)^{\beta}},$$

где α и β – параметры алгоритма, а $\mathrm{Est}(r_2 r_3 \ldots r_{i-1}, dir_j)$ – априорная оценка. Для ее вычисления предполагается, что следующая аминокислота располагается по направлению dir_j и подсчитывается количество гидрофобных (n_H) и полярных (n_P) остатков в соседних с ней узлах. Чтобы образовывались новые связи, гидрофобные остатки целесообразно располагать рядом с гидрофобными или со свободными узлами (есть шанс заполнить их гидрофобными остатками при дальнейшем построении молекулы), а полярные не ставить рядом с гидрофобными (оставляя место для возможных новых соединений). Такая интерпретация

аналогична естественному поведению молекулы: полярные остатки стремятся к соседству с полярным окружением – водой или другими полярными остатками.

Введем параметры $0 \le est_{HP} < est_{H0} < est_{HH}$, $est_{HP} + est_{H0} + est_{HH} = 1$, $0 \le est_{PH} < est_{P0} \le est_{PP}$, $est_{PH} + est_{P0} + est_{PP} = 1$, и определим оценку $\mathrm{Est}(r_2 r_3 ... r_{i-1}, dir_j)$ следующим образом:

$$\mathrm{Est}(r_2 r_3 ... r_{i-1}, dir_j) = \begin{cases} n_P * est_{HP} + n_H * est_{HH} + (n - n_P - n_H) * est_{H0}, & o_{i+1} = H, \\ n_P * est_{PP} + n_H * est_{PH} + (n - n_P - n_H) * est_{P0}, & o_{i+1} = P. \end{cases}$$

Приведенная оценка является обобщением известных ранее – в литературе часто встречается комбинация $est_{PH} = est_{P0} = est_{PP} = 1/3$, $est_{H0} = est_{HP} = 0$, $est_{HH} = 1$. Ниже для вычислительного эксперимента были выбраны значения $est_{PH} = 0.2$, $est_{P0} = est_{PP} = 0.4$, $est_{HP} = 0.15$, $est_{H0} = 0.25$, $est_{HH} = 0.6$.

Чтобы новые данные учитывались как более актуальными, используется процедура *ИспаритьФеромоннуюМатрицу*(Φ), в которой каждый элемент матрицы умножается на параметр $0 \le \rho \le 1$, задающий степень испарения феромонов. Процедура *ОбновитьФеромоннуюМатрицу*(Φ, v) в ее часто встречающейся реализации добавляет к каждому элементу, который соответствует направлению, содержащемся в структуре $v = r_2 r_3 ... r_{N-1}$, значение, пропорциональное энергии v:

$$\Phi[k_{i+1}, i] = \Phi[k_{i+1}, i] + (-E(v))^{\gamma},\ i = \overline{1, N-2},$$

где k_i определяется условием $dir_{k_i} = r_i$, а $\gamma > 0$ – параметр алгоритма.

С учетом особенностей задачи предлагается альтернативный вариант обновления феромонной матрицы. Его идея состоит в том, чтобы усиливать только те феромонные дорожки, которые влияют на энергию заданной молекулы. Для этого используется следующая процедура: введем массив ϕ длиной $N-2$ и заполним его нулями. Пусть l, m – номера остатков, образующих гидрофобную связь, $l < m$. В образовании этой связи принимают участие элементы $r_l, r_{l+1}, ..., r_{m-1}$. Подкорректируем массив ϕ для этой связи так:

$$\phi[i] = \phi[i] + 1,\ i = \overline{\max\{1, l-1\},\ m}$$

и повторим эту операцию для всех гидрофобных связей в структуре $v = r_2 r_3 ... r_{N-1}$, после чего обновим феромонную матрицу:

$$\Phi[k_{i+1}, i] = \Phi[k_{i+1}, i] + \phi[i]^{\gamma},\ i = \overline{1, N-2}.$$

Такая схема позволяет более точно учитывать приемлемость той или иной части структуры молекулы.

4. Вычислительный эксперимент. Для сравнения и проверки эффективности разработанных алгоритмов был проведен вычислительный эксперимент. Из базы данных протеинов с известной структурой [7] было выбрано 11 примеров

размерностью от 52 до 302 аминокислот. Рассматривалось 8 вариантов алгоритма – с абсолютным и относительным кодированием структуры, без априорных оценок и с использованием вышеизложенного метода для их подсчета, с каждым из двух описанных способов обновления феромонной матрицы. Локальный поиск не применялся. Учитывая стохастичность алгоритма, для каждой задачи было выполнено три рестарта. Предложенные модификации алгоритма доказали свою целесообразность как в случае с абсолютной кодировкой, так и с относительной. Сравнительные результаты работы четырех вариантов алгоритма на базе относительной кодировки показаны на рис. 2.

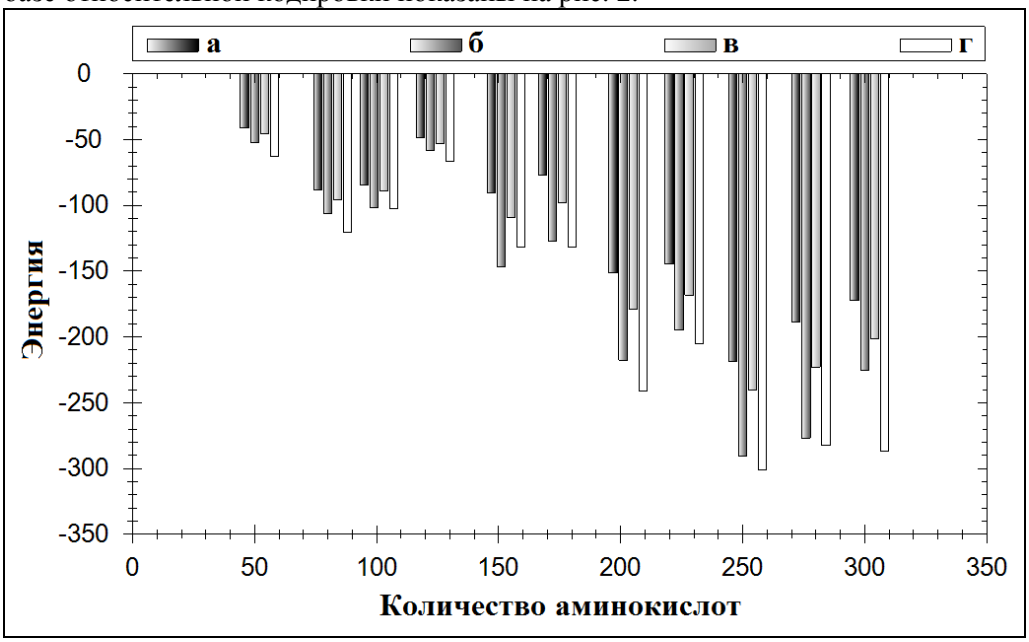

Рис. 2. Результаты вычислительного эксперимента

На диаграмме модификации ***а*** и ***б*** не используют априорных оценок на этапе построения структуры, в отличие от ***в*** и ***г***. Модификации ***б*** и ***г*** обновляют феромонную матрицу первым из вышеописанных способов, ***а*** и ***в*** – вторым. По оси X отложено количество аминокислот в исследуемой молекуле, по оси Y – усредненная (по трем запускам алгоритма) энергия оптимальной свертки, которая является целевой функцией задачи.

Выводы. Экспериментальные исследования подтверждают, что использование априорных оценок и обновление феромонной матрицы по частям существенно улучшают эффективность алгоритма метода ОМК по сравнению с вариантами без использования этих процедур. Открытыми остаются вопросы эффективной настройки параметров алгоритма и схема внедрения процедуры локального поиска для нахождения более оптимальных решений.

В.О. Рудик

АНАЛІЗ ЕФЕКТИВНОСТІ АЛГОРИТМІВ ОМК ДЛЯ РОЗВ'ЯЗАННЯ ЗАДАЧІ ПРО ЗГОРТАННЯ ПРОТЕЇНУ

Розглядається гідрофобно-полярна модель згортання протеїну, пропонується та досліджується алгоритм прогнозування третинної структури білка, розроблений на базі методу ОМК. Будується модель молекули протеїну в тривимірній трикутній решітці. Запропоновані методи апріорної оцінки позиції в решітці та поновлення матриці феромонних шляхів, їх ефективність підтверджена обчислювальним експериментом.

V.O. Rudyk

ANT COLONY OPTIMIZATION ALGORITHMS FOR PROTEIN FOLDING PROBLEM ANALYSIS

The Hydrophobic-Hydrophilic protein folding model is examined, the algorithm based on ACO optimization method is proposed and analyzed for protein tertiary structure prediction. The molecule model is considered to be a chain whose monomers are placed on the vertices of 3D triangular lattice. The a priori position estimation method and pheromone trails updating method are deviced, computational experiment confirms their appropriateness.

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. *Shmygelska A, Hoos H.* An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem // BMC Bioinformatics. – 2005. – N 6(30). – P.30–52.
3. *Local* rules for protein folding on a triangular lattice and generalized hydrophobicity in the HP model / *Agarwala R., Batzoglou S., Dancik V* et al. // J. of Computational Biology. – 1997. – N 4. – P. 275–296.
4. *Fidanova S., Lirkov I.* Ant Colony System Approach for Protein Folding // Int. Conf. Multiconference on Computer Science and Information Technology. – 2008. – P.887–891.
5. *Hart W., Istrail S.* Lattice and Off-Lattice Side Chain Models of Protein Folding: Linear Time Structure Prediction Better Than 86% of Optimal // J. of Computational Biology.– 1997. – N 4. – P.241-259.
6. *Рудык В.* Представление структуры белка в трехмерных дискретных решетках произвольного типа // Теорія оптимальних рішень. – 2011. – №10. – С. 38–47.
7. Information Portal to Biological Macromolecular Structures: http://www.rcsb.org/

Получено 03.04.2012