УДК 519.21

Л.Ф. ГУЛЯНИЦКИЙ, В.А. РУДЫК

МОДЕЛИРОВАНИЕ СВЕРТЫВАНИЯ ПРОТЕИНА В ПРОСТРАНСТВЕ

Рассматривается проблема прогнозирования третичной структуры протеина по заданной последовательности аминокислот. На основе HP-модели она формализуется в виде специальной задачи комбинаторной оптимизации, определенной на трехмерной треугольной решетке. Предложены два алгоритма локального поиска, эффективность которых исследована путем анализа результатов проведенного вычислительного эксперимента.

Введение. Задача определения пространственной структуры протеина по последовательности аминокислотных остатков – весьма важная проблема в современной биохимии, молекулярной биологии и биофизике [1]. Экспериментальное определение структуры белка не всегда возможно и целесообразно – учитывая сложность, высокую стоимость и ограниченность возможностей экспериментальных методик. Этих проблем можно избежать, прогнозируя форму молекулы математическими методами, основанными на физических и эмпирических предпосылках. Структура протеина может быть описана на разных уровнях детализации. Исследователями разработан ряд математических моделей проблемы свертывания протеина, которые абстрагируются от тех или иных факторов, концентрируясь на важнейших характеристиках. Весьма представительный класс образуют модели, использующие плоские или пространственные решетки, которые приводят к задачам комбинаторной оптимизации [2].

Для решения возникающих задач комбинаторной оптимизации (ЗКО) были предложены точные алгоритмы, прежде всего метода ветвей и границ, предназначенные для решения задач на плоскости. Однако трудоемкость таких алгоритмов и приближенный характер модели делают нецелесообразным, а зачастую – и невозможным их применение для прогнозирования структуры молекул длины больше 50. Этим объясняется появление все большего числа приближенных алгоритмов – от траекторных алгоритмов до гибридных метаэвристик [3–8].

В работе используется более адекватная, хотя и более сложная с вычислительной точки зрения трехмерная HP-модель. Большинство известных моделей этого типа основано на кубических решетках [4–6]. Цель данного исследования – разработка подходов к прогнозированию структуры протеина на основе использования трехмерной треугольной решетки, которые продолжают и обобщают результаты из [7]. Для решения возникающей ЗКО разработаны два алгоритма локального поиска. Их эффективность иллюстрируется результатами вычислительного эксперимента по прогнозированию молекул длины от 500 до 1500. В заключение обсуждаются направления возможных дальнейших исследований.

1. HP-модель Дилла

Задача определения третичной структуры протеина состоит в прогнозировании формы молекулы в пространстве исходя из последовательности аминокислотных остатков в ее цепи. Одна из самых распространенных моделей, описывающая этот процесс – гидрофобно-полярная модель (HP-модель), впервые предложенная Диллом [9]. В соответствии с ней каждая из 20 аминокислот относится к одному из двух классов – гидрофобных (H) или полярных (P), а аминокислотную последовательность можно рассматривать как последовательность $S = (s_1, s_2, ..., s_n)$, $s_i \in \{H, P\}$, $i = \overline{1,n}$, длины n. Каждый элемент s_i располагается в узле некоторой решетки так, чтоб соседние в последовательности элементы соответствовали соседним узлам решетки, таким образом определяется путь в решетке, называемый сверткой. Если этот путь не содержит самопересечений, свертка считается допустимой.

Согласно терминологии Дилла, связанными соседями называются остатки, которые являются соседними в последовательности, а топологическими соседями – те, которые расположены в соседних узлах решетки и не являются связанными соседями. Пары топологических соседей H-H образуют связь. Энергия свертки подсчитывается как количество связей в ней со знаком минус.

Более формально, если каждый элемент s_i, $i = \overline{1,n}$, аминокислотной последовательности отображается в узел прямоугольной решетки $L_i = (x_i, y_i)$, то энергия $E(S)$ может быть определена так:

$$E(S) = - \sum_{1 \le i \le j-2 \le n-2} I(L_i, L_j)\, h(s_i, s_j) \ ,$$

где $$I(L_i, L_j) = \begin{cases} 1, \text{ если узел } L_i \text{ является соседним к узлу } L_j, \\ 0 - \text{в противном случае,} \end{cases}$$

$$h(s_i, s_j) = \begin{cases} 1, \text{ если } s_i \text{ и } s_j \text{ являются гидрофобными,} \\ 0 - \text{в противном случае.} \end{cases}$$

Аналогично определяется величина энергии и в случае трехмерной кубической решетки.

Задача определения третичной структуры протеина сводится к поиску свертки с наименьшей энергией. Доказано, что в такой постановке задача является NP-сложной [10, 11].

В работе рассмотрена менее исследованная специальная трехмерная модель протеинов как более соответствующая реальным пространственным молекулам и учитывающая многие проблемы, возникающие при переходе от упрощенной 2D модели к 3D.

2. Пространственная модель

Простейшая трехмерная решетка – кубическая – в контексте поставленной задачи обладает тем свойством, что два остатка могут оказаться соседями по решетке тогда и только тогда, когда число элементов между ними в аминокислотной последовательности будет четным. Это не соответствует логическим представлениям, так как, например, строка $(HP)^n$ не будет иметь ни одной связи. Поэтому для исследования была выбрана трехмерная треугольная решетка (рис. 1) [8] – в ней для любых двух остатков существует свертка, в которой они будут располагаться в соседних узлах решетки.

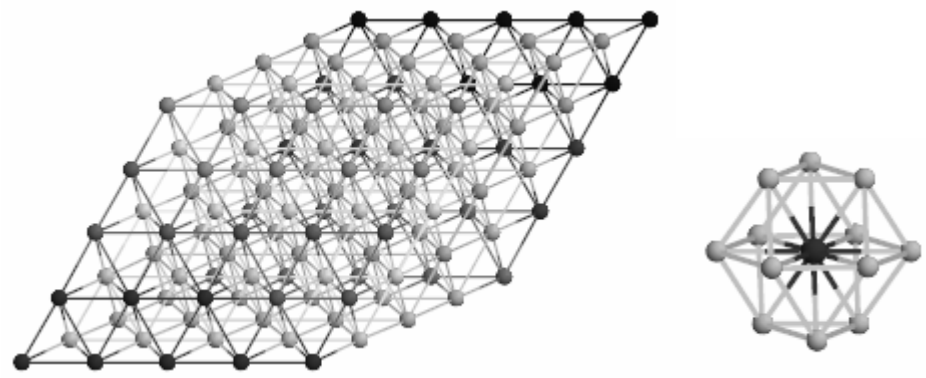

РИС. 1. Трехмерная треугольная решетка

В выбранной решетке вводится система целочисленных координат по следующему правилу: узлу решетки с координатами (x, y, z), $x, y, z \in Z$, Z – множество целых чисел, ставится в соответствие пространственная точка $xe_1 + ye_2 + ze_3$, где $e_1 = \left(\frac{\sqrt{3}}{2}, \frac{1}{2}, 0\right)$, $e_2 = (0, 1, 0)$, $e_3 = \left(\frac{\sqrt{3}}{6}, \frac{1}{2}, \frac{\sqrt{33}}{6}\right)$.

Узлы с координатами (x_1, y_1, z_1) и (x_2, y_2, z_2) будут соседними тогда и только тогда, когда $(x_1 - x_2, y_1 - y_2, z_1 - z_2) \in U$, где U – множество направлений сдвига

$$U = \{(0,1,0), (1,0,0), (1,-1,0), (-1,0,0), (-1,1,0), (0,-1,0),$$
$$(0,0,1), (0,-1,1), (-1,0,1), (0,1,-1), (1,0,-1), (0,0,-1)\}.$$

В соответствии с введенными терминами свертку можно представить в виде последовательности точек с целочисленными координатами $p_1 p_2 ... p_n$, $p_i = (x_i, y_i, z_i)$, $x_i, y_i, z_i \in Z$, $i = \overline{1,n}$, – такая кодировка называется координатной. Однако в ней есть ряд недостатков:

– в общем виде такая последовательность не гарантирует, что связанные соседи будут соседями в решетке;

– свертка может быть задана неоднозначно – параллельный перенос и поворот приведут к разным представлениям, что расширяет область поиска решений.

Для предотвращения этих недостатков предлагается использовать другую кодировку, называемую абсолютной. В ней свертка представляется в виде последовательности векторов направлений $a_1 a_2 ... a_{n-1}$, $a_i \in U$, $i = \overline{1,n-1}$. Для перехода к координатной кодировке достаточно воспользоваться выражениями $p_1 = (0,0,0)$, $p_i = p_{i-1} + a_{i-1}$, $i = \overline{2,n}$.

В таком представлении каждая последовательность задает определенный путь в решетке, а для проверки его допустимости необходимо только проверить на самопересечения. Проблема с неоднозначностью решается только частично – при повороте свертки ее кодировка меняется.

Альтернативой абсолютной кодировке может быть относительная кодировка, в которой вектор направления зависит от части предыдущей свертки. Если для двухмерной решетки в абсолютной кодировке есть шесть направлений – «Запад», «Северо-запад», «Северо-восток», «Восток», «Юго-восток» и «Юго-запад», то в относительной кодировке их уже пять – «Назад-налево», «Вперед-налево», «Вперед», «Вперед-направо», «Назад-направо» (рис. 2).

Преимущества такого представления следующие:

– каждая свертка однозначно задается своей относительной кодировкой;

– отсутствием направления «назад» частично решается проблема самопересечения;

– изменение отдельного элемента последовательности в относительной кодировке приводит к повороту части свертки, которая находится после него, в то время как в абсолютной кодировке – только к параллельному переносу. Это позволяет получать более значительные изменения энергии при малом изменении последовательности.

Если в двухмерной решетке вектор поворота по определенному относительному направлению зависит только от предыдущего абсолютного направления в этой свертке (так, если последнее абсолютное направление было «Восток»,

то следующее относительное направление «Вперед-направо» будет соответствовать абсолютному «Юго-восток»), то в трехмерной – от предыдущих двух. Для перевода абсолютной кодировки в относительную и наоборот строится таблица, первые два столбца которой заполняются возможными абсолютными направлениями, третий – относительным, четвертый – соответственным абсолютным.

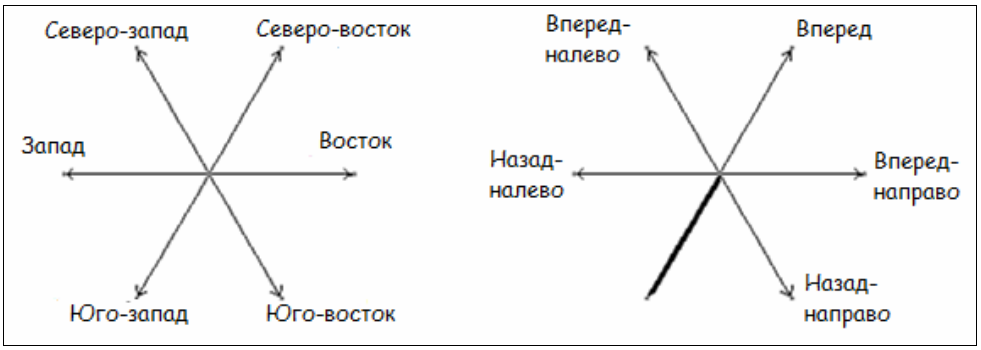

РИС. 2. Направления поворота в абсолютной и относительной кодировке для двухмерной треугольной решетки

При использовании относительной кодировки пространством решений задачи с входящей последовательностью длины n будет множество

$$X = \{r_1 r_2 ... r_{n-2},\ r_1 \in \tilde{U}, r_i \in U \setminus \{(0,-1,0)\}, i = \overline{2, n-2}\},$$

где $\tilde{U} = \{(0,1,0), (1,0,0), (1,-1,0), (1,0,-1)\}$, с метрикой

$$\rho(v_1 v_2 ... v_{n-2},\ w_1 w_2 ... w_{n-2}) = \sum_{i=1}^{n-2} (1 - \chi(v_i, w_i)),$$

$$\chi(v_i, w_i) = \begin{cases} 1, v_i = w_i \\ 0, v_i \neq w_i \end{cases}.$$

В контексте таких обозначений множество U задает не направление в решетке, а взаимное расположение векторов направлений относительно друг друга. Множество $\tilde{U}$ задает все возможные взаимные расположения первых трех элементов в предположении, что первые два элемента находятся в узлах $(0,0,0)$ и $(0,1,0)$ соответственно. Следовательно, положение свертки в пространстве четко фиксировано. Множество допустимых решений задачи $D \subset X$ – это множество всех сверток без пересечений. Таким образом, задача прогнозирования третичной структуры молекулы протеина состоит в поиске

$$x^* = \arg \min_{x \in D \subseteq X} E(x),$$

где $E(x)$ – энергия свертки.

3. Детерминированный локальный поиск

Для минимизации энергии протеина в построенной модели предлагается алгоритм, относящийся к детерминированному локальному поиску [12]. Сам по себе он может и не найти качественного решения, поскольку функция энергии многоэкстремальна, но из-за относительно небольшого затрачиваемого времени может быть использован в более сложных алгоритмах или при решении задач повышенной размерности.

Для произвольного $v = v_1 v_2 ... v_{n-2} \in X$ обозначим:

$$U_i = \begin{cases} \tilde{U}, & i = 1, \\ U \setminus \{(0,-1,0)\}, & i = \overline{2, n-2}, \end{cases}$$

$$v[i] = v_i .$$

Алгоритм локального поиска, использующий окрестности минимального радиуса, показан на рис. 3. Здесь «mod» обозначает деление по модулю.

```
procedure LocalSearch ( s0 )
    lastIndxChanged = 1;
    indx = 1;
    repeat
        v := s0 ;
        foreach r из U_indx do
            v[indx] := r ;
            if v ∈ D & E(v) < E(s0) then
                s0 := v ;
                lastIndxChanged:= indx;
            end if;
        end foreach;
        indx =(indx mod (n-2))+1;
    until lastIndxChanged = indx;
return s0 ;
end procedure.
```

РИС. 3. Схема алгоритма детерминированного локального поиска с окрестностью радиуса 1

При необходимости получения более точных решений можно нарастить область поиска на каждом шаге, увеличив радиус поиска (схема такого алгоритма представлена на рис. 4). Естественно, что при применении усовершенствованного алгоритма увеличивается и время решения задачи – более детальное исследование разработанных алгоритмов оптимизации приведено далее.

```
procedure LocalSearch2 ( s0 )
    lastIndxChanged = 1;
    indx = 1;
    repeat
        v := s0 ;
        foreach r1 из U_indx do
            v[indx] := r1;
            foreach r2 из U_indx+1 do
                v[indx+1] := r2;
                if v ∈ D & E(v) < E(s0) then
                    s0 := v;
                    lastIndxChanged:= indx;
                end if;
            end foreach;
        end foreach;
        indx =(indx mod (n–3))+1;
    until lastIndxChanged = indx;
return s0 ;
end procedure.
```

РИС. 4. Схема алгоритма детерминированного локального поиска с окрестностью радиуса 2

4. Вычислительный эксперимент

Для анализа эффективности предложенных алгоритмов был проведен вычислительный эксперимент, результаты которого приводятся в таблице.

Задача решалась для случайно сгенерированных молекул длины 500–1500 *c* различными соотношениями гидрофобных и полярных остатков, для каждой размерности и соотношения решалось по 3 задачи, всего 60 задач. В таблице n – длина входной последовательности; H:P – соотношение гидрофобных остатков к полярным; $\overline{E_0}$ – значение усредненной (по трем задачам) энергии для начальной свертки; $\overline{t_1}, \overline{t_2}$ – усредненное время в секундах, потраченное алгоритмами локального поиска с окрестностью радиуса 1 и 2 соответственно; $\overline{E_1}$, $\overline{E_2}$ – усредненные значения энергии, найденные алгоритмами; $\overline{q_1}$, $\overline{q_2}$ – усредненные значения улучшения: $q_i = \dfrac{E_0 - E_i}{|E_0|} \cdot 100\ \%,\ i = 1, 2.$

ТАБЛИЦА

n	H:P	$\overline{E_0}$	$\overline{t_1}$	$\overline{E_1}$	$\overline{q_1}$	$\overline{t_2}$	$\overline{E_2}$	$\overline{q_2}$
500	4:1	–331	37136,7	–690,3	108,8	308934,3	–778,7	136,1
500	2:1	–219,3	28003,3	–504,3	130,7	211032	–627,7	188,3
500	1:1	–125,7	36915,7	–358,7	190,3	205183,7	–395,7	222,5
500	1:2	–57	30199,7	–178,7	217,4	213269,3	–229,7	308,9
500	1:4	–26	24940,3	–79	211,7	162412,3	–120,7	381,3
1000	4:1	–632,7	152917,3	–1271,7	100,9	940816,3	–1492,7	135,9
1000	2:1	–377,7	113836,3	–913,7	143	967104,3	–1227,3	227,5
1000	1:1	–231,3	93396	–537,7	132,2	692723,3	–712	207,5
1000	1:2	–115,7	74764,3	–305,7	169,4	481854	–493,7	336,7
1000	1:4	–40	126953	–152	280,9	595179	–216	441,7
1500	4:1	–999,7	314950,3	–1822,3	82,3	2202480	–2262	126,1
1500	2:1	–621,3	272091,7	–1276,7	105,6	2409250	–1521	145
1500	1:1	–383,3	370674	–905,3	137,7	1990843	–1060,3	177,9
1500	1:2	–164,7	285044,3	–453	176,9	1683528	–652	298
1500	1:4	–65,7	230893	–208,3	219,9	1458758	–317,7	396,1

На рис. 5 приведен пример полученной свертки молекулы с 1000 аминокислотными остатками. Светлые шары соответствуют гидрофобным, а темные – полярным остаткам.

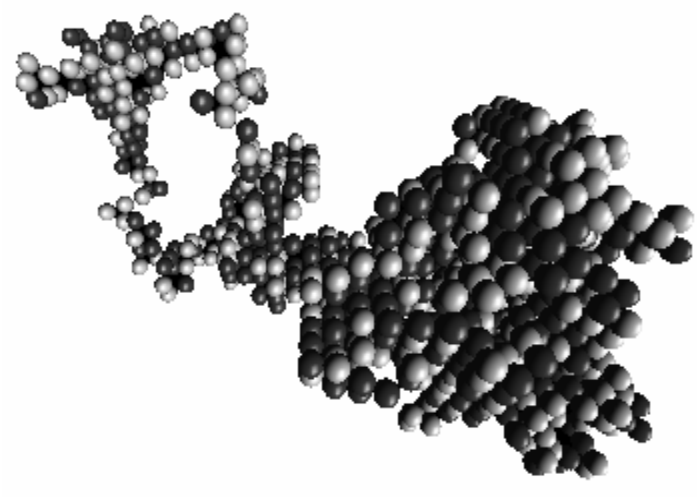

РИС. 5. Результат оптимизации свертки молекулы (n = 1000)

Заключение. Для моделирования процесса сворачивания протеина использована и исследована трехмерная треугольная сетка. На примере процедур локального поиска показана возможность использования относительной кодировки в алгоритмах оптимизации. Разработанные алгоритмы локального поиска могут быть использованы как в качестве блоков при построении более сложных (метаэвристических) методов оптимизации, так и для решения задач повышенной размерности.

Целью дальнейших исследований может быть разработка и применение алгоритмов стохастического локального поиска и других метаэвристических алгоритмов на основе приведенных способов кодировки структуры молекулы.

Л.Ф. Гуляницький, В.О. Рудик

МОДЕЛЮВАННЯ ЗГОРТАННЯ ПРОТЕЇНУ У ПРОСТОРІ

Розглядається проблема прогнозування третинної структури протеїну за заданою послідовністю амінокислот. На основі HP-моделі вона формалізується у вигляді спеціальної задачі комбінаторної оптимізації, яка визначена на тривимірній трикутній решітці. Запропоновано два методи локального пошуку, ефективність яких досліджена шляхом аналізу результатів проведеного обчислювального експерименту.

L.F. Hulianytskyi, V.A. Rudyk

3-D MODELLING OF PROTEIN FOLDING

The problem of protein tertiary structure prediction from its amino acid sequence is examined. Basing on HP-model, it is formalized as a specific combinatorial optimization problem defined on a three-dimensional triangular lattice. Two local search methods are proposed and their efficiency is examined by analyzing the results of computational experiment.

1. *Гупал А.М., Сергиенко И.В.* Оптимальные процедуры распознавания. – К.: Наук. думка, 2008. – 232 с.
2. *Greenberg H.J., Hart W.E., Lancia G.* Opportunities for Combinatorial Optimization in Computational Biology // Informs J. on Computing. – 2004. –**16**, N 3. – P. 211–231.
3. *Agarwala R., Batzoglou S., Dancik V., Decatur S., Farach M., HannenhallI S., Muthukrishnan S., Skiena S.* Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the HP model // J. of Computational Biology. –1997. – **4**, N 3. – P. 75–96.
4. *Hart W., Istrail S.* Fast protein folding in the hydrophobic-hydrophilic model within three-eighths of optimal // J. of Computational Biology. – 1996 –**3**(1). –P. 53–96.
5. *Thachuk C., Shmygelska A., Hoos H.* A replica exchange Monte Carlo algorithm for protein folding in the HP model // BMC Bioinformatics. – 2007. – **8**. – P. 342–361.
6. *Thalheim T., Merkle D., Middendorf M.* Protein Folding in the HP-Model Solved With a Hybrid Population Based ACO Algorithm // Int. J. of Computer Science. – 2006. – **35**. – P. 3–12.
7. *Гуляницкий Л.Ф., Рудык В.А.* Разработка и исследование алгоритмов решения задачи прогнозирования третичной структуры протеина / Intelligent Support of Decision Making (Eds. Krassimir Markov et al.). Int. Book Series "Information science and computing". N 10. – Sofia: ITHEA, 2009. – P. 97 – 103.

8. *Decatur S.* Protein Foldins in the generalized hydrophobic-polar model on the triangular lattice // Technical Rep. MIT-LCS-TM-559. – Massachusetts Institute of Technology, May 1998. – 9 p.
9. *Dill K., Bromberg S., Yue K., Fiebig K.M., Yee D., Thomas P., Chan H.* Principles of protein folding – a perspective from simple exact models // Protein Science. – 1995. – **4**. – P. 561– 602.
10. *Berger B., Leighton T.* Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete // J. of Computational Biology. – 1998. – **5**(1). – P. 27–40.
11. *Crescenzi P., Goldman D., Papadimitriou C., Piccolboni A., Yannakakis M.* On the complexity of protein folding // J. of Computational Biology. –1998. – **5**(3). – P. 423–465.
12. *Сергиенко И.В.* Математические модели и методы решения задач дискретной оптимизации. – К.: Наук. думка, 1988. – 472 с.

Получено 27.11.2009

Об авторах:

Гуляницкий Леонид Федорович,
доктор технических наук, ведущий научный сотрудник
Института кибернетики имени В.М. Глушкова НАН Украины,
leonhul.icyb@gmail.com

Рудык Виталина Александровна,
студентка Киевского национального университета им. Тараса Шевченко.
vitalina.rudyk@gmail.com